Attachment A

Application Cover Sheet

Name of Respondent Slalom			
Mailing Address 255 S King St	City/Town Seattle	State WA	Zip Code 98104
Telephone (206) 438-5700	Fax N/A	Web Address www.slalom.com	
Primary Contact for Clarification Russell Norris		Primary Contact E-mail Address russelln@slalom.com	
Authorized Signatory Russell Norris		Authorized Signatory E-mail Address russelln@slalom.com	
Legal Status/Jurisdiction(e.g., a Massachusetts Corporation, LLC, LLP, etc.)		Respondents W-9 Respondents UEI No. 84-1246887	

07 OCTOBER 2025

AI Hub Data Commons Collaborative Consulting Services

Slalom's Application to Massachusetts Technology Collaborative RFP No. 2026-JAII-01

Slalom, Inc.
Office Address: 399 Boylston Street Suite 1000, Boston, MA 02116
Main Phone: (617) 316-5400

slalom

Massachusetts Technology Collaborative
75 North Drive
Westborough, MA 01581-3340

To Sabrina Mansur, Procurement Team Leaders, and Massachusetts Technology Collaborative Stakeholders:

On behalf of Slalom, Inc., I am pleased to submit our response to RFP No. 2026-JAII-01 for the AI Hub Data Commons Collaborative Consulting Services. We are excited by the vision of the Massachusetts AI Hub and the Data Commons Collaborative (DCC), and we share your commitment to building a secure, ethical, and extensive platform that empowers innovation across the Commonwealth.

Slalom's proposal addresses your requirements for a metadata-rich, searchable catalog aligned with FAIR principles, synthetic data capabilities, and embedded AI fairness and bias detection tools, robust governance and security, and accessibility for diverse communities. Our phased approach brings rapid delivery of value, sustainable budgeting, and clear program management, with work commencing in December 2025.

We believe Slalom is uniquely positioned to deliver on these needs. Our approach blends technical expertise in scaled solutions involving cloud-native architecture, data engineering and governance, and AI solution deployment, including responsible AI frameworks. Our approach is characterized by:

- **Broad, Deep, and Customizable Services:** Slalom offers a comprehensive suite of consulting services spanning strategy, data, artificial intelligence, cloud, systems implementation, experience strategy & design, digital product building, organizational change, and more. Our teams combine technical excellence with human-centered design to deliver solutions that are practical, impactful, and tailored to each client's unique environment.
- **Industry-Leading Partnerships:** We partner with over 700 of the world's leading technology providers—including AWS, Databricks, and Snowflake. These alliances give our clients priority access to vendor resources, early insights into product roadmaps, and proven integration patterns.
- **Local Commitment, Global Reach:** Our Boston-based team of over 400 consultants, designers, engineers, and data scientists brings deep expertise and local context, backed by the resources and experience of a global firm.
- **Agile, Collaborative Culture:** Slalom's approach is fiercely human. We work side-by-side with our clients, embedding ourselves within their teams to foster true partnership, transparency, and knowledge transfer. Our agile methodologies and employee-centric culture empower us to respond quickly to changing needs and deliver measurable results.

We are committed to partnering with the Massachusetts Technology Collaborative (MassTech) and the AI Hub to launch the DCC, positioning Massachusetts as a global leader in ethical, inclusive, and innovative AI. We are committed to a collaborative partnership and welcome discussion regarding the overall cost and scope of this engagement. Please feel free to contact me directly at RussellN@Slalom.com with any questions.

Thank you,
Russell Norris, Senior Managing Director, Slalom New England

slalom

Respondent Overview **5**
Proposed Approach to Data Commons **11**
Phasing Strategy **15**
Solution Summary **17**
Team Structure **24**
Collaboration & Delivery Model **26**
Design Approach **31**
Technology Approach **35**
Adoption Approach **43**
Organizational Enablement Approach **46**
Key Topic: Data Catalog **48**
Key Topic: Commons Credits System **52**
Key Topic: Fairness Tooling **57**
Key Topic: Synthetic Data Generation **61**
Key Topic: User-Contributed Data **68**
Appendix - Team & Staffing **70**
Budget & Pricing (Attachment C) **90**
Appendix – RFP Milestones **91**
Appendix – Customer References **92**
Appendix – Team Structure **97**
Appendix – Customer Stories **99**
W-9 **108**

Table of Figures

Figure 1: Slalom Services 8
Figure 2: Project Timeline Overview 12
Figure 3: Project Timeline With Risks 17
Figure 4: User Interactions and System Architecture 18
Figure 5: Technical Architecture 19
Figure 6: Proposed Full Team Structure 26
Figure 7: Product Engineering Methodology Process 28
Figure 8: The "Big 3" for Delivery Success 29
Figure 9: Sample Status Report 30
Figure 10: Project Timeline With Design Phases 32

Figure 13: Project Timeline With Adoption Phases 44
Figure 14: Illustrative Discovery & Design Phase Activities 45
Figure 15: Illustrative Build Phase Activities 46
Figure 16: Example of a Mission Model Canvas 47
Figure 17: Project Timelines With Enablement Activities 48
Figure 18: Project Timelines With Credit System Phases 55
Figure 19: Illustrative Ruleset for Testing in MVP Based on Behavioral Hypotheses 56
Figure 20: Example Scenarios 59
Figure 21: User-Facing Synthetic Data Generation Process 62
Figure 22: Synthetic Data Studio User & System Overview 63
Figure 23: Illustrative Synthetic Data Generation Feature Timeline 66
Figure 24: RFP Milestones 91
Figure 25: Team Structure 98

Respondent Overview

Our Understanding

Massachusetts is uniquely positioned to lead in artificial intelligence (AI) innovation, with its world-class academic institutions, healthcare systems, and technology ecosystem. This potential is constrained by critical structural and operational gaps in the state's data landscape. At the heart of this challenge is the absence of a unified, trusted, and equitable framework for data sharing, governance, and infrastructure—one that enables high-quality, representative datasets to be accessed and used responsibly across the Commonwealth.

Data without borders.

Massachusetts has the talent and ambition to lead in AI, but its fragmented data ecosystem is holding it back. Without a unified framework for data sharing and governance, innovation is stalling, and equity is slipping—especially for smaller players. The MassTech Collaborative Data Commons Collaborative offers not just a solution, but a future-ready foundation for responsible, inclusive AI leadership.

Unlocking this potential is not just a technical imperative—it is a strategic opportunity to drive inclusive growth and public value. A modern data ecosystem would empower a wide range of stakeholders—including independent researchers, nonprofits, startups, public agencies, and higher education institutions—to develop solutions that improve transportation planning, enhance healthcare delivery, support sustainable agriculture, advance academic research, and more. The federal government and other states have demonstrated the power of publicly available data—from applications that use location data for real-time noise analysis in U.S. Department of Housing and Urban Development (HUD) housing projects, to tools that streamline water data for real estate agents, to logistics platforms that help small farmers optimize truck space. Massachusetts is uniquely positioned to lead in this space.

Today, however, data silos persist across public, academic, and private institutions. These silos:

- Impede interoperability and reuse,
- Duplicate data preparation efforts, and
- Slow the AI development lifecycle from discovery to deployment.

The lack of common metadata standards, technical schemas, and legal frameworks further inhibits integration and reuse of data across domains. This not only slows innovation but also increases the cost and complexity of AI development.

Governance gaps compound these issues. Without standardized protocols for data stewardship, consent management, and ethical oversight, organizations face uncertainty around compliance with state and federal regulations and emerging norms related to fairness, transparency, and algorithmic accountability. These challenges are most acute in sensitive sectors such as healthcare, education, and public infrastructure.

Without decisive action, Massachusetts risks falling behind other states and countries that are investing in shared data ecosystems, regulatory sandboxes, and AI-enabling infrastructure. A fragmented data landscape not only slows innovation but also erodes public trust in systems built on biases or opaque data.

Slalom understands that the MassTech Data Commons Collaborative (DCC) is more than a technical solution—it is a foundational investment in the Commonwealth's future. The success of the investment requires a strong partner like Slalom who has delivered public-facing Data, AI, and Cloud solutions at scale, bringing deep expertise in the following areas relevant for the DCC:

- Human-Centered Design,
- Scaled Data Architecture and Governance,
- Cloud-native Architecture,
- Cutting-edge, Scaled AI solutions and Fairness Frameworks,
- Accelerated Engineering, and
- Organizational Transformation.

We are honored to submit our response to RFP No. 2026-JAII-01 and to partner with MassTech in delivering this innovative program.

Company Overview & Qualifications

Slalom is a purpose-driven consulting firm with more than 10,000 professionals across 52 offices globally including a strong local presence in Boston and greater New England. We

partner with organizations to deliver strategy, technology, and transformation solutions that drive meaningful impact. Our Boston office, established in 2012, is home to over 300 consultants, designers, engineers, and data scientists who are deeply invested in the success of the Commonwealth and its communities.

Who we are.

We believe consulting should be personal, agile, and tailored. Our locally empowered teams build lasting relationships and deliver solutions grounded in each client's unique environment. Because our consultants live and work in the communities we serve, our work reflects local context and stakeholder priorities.

Slalom services.

Slalom brings together a unique combination of Cloud and Data expertise, large program delivery excellence, strong Alliance partnerships with a local presence in Massachusetts to reduce project risk and accelerate time-to-value—achieving meaningful, lasting outcomes for Massachusetts.

- **Local Commitment, Global Reach |** Our Boston-based team brings direct experience working with Massachusetts agencies, healthcare institutions, and higher education entities. We pair this local knowledge and context with the resources and expertise of a global firm.
- **End-to-end Capabilities |** From strategy and design to implementation and change management, Slalom delivers holistic solutions that address the full lifecycle of technology and data initiatives. Our transformation capability helps organizations rethink how they deliver value by aligning people, processes, and technology to drive innovation.
- **Innovation with Purpose |** We are recognized leaders in scaled cloud-native architecture, data architecture, engineering, and governance, AI and synthetic data tooling, and responsible AI. Our teams have implemented solutions using leading open-source and commercial tools, and we are committed to advancing ethical, responsible AI.
- **Collaborative Partnerships |** Our strategic alliance partnerships with leading technology providers—including AWS, Microsoft, Google Cloud, Salesforce, Databricks, and Snowflake—position Slalom to bring priority access to vendor resources, early insights into product roadmaps, and proven integration patterns.

This unique combination is supported through Slalom's national services. A comprehensive view of our services can be found in Figure 1: Slalom Services.

Slalom services
Slalom's services are broad, deep, and infinitely customizable. From strategy through delivery, we collaborate with you to bring powerful customer experiences, innovative ways of working, and new products, services, and businesses to life.
St Strategy
• Business strategy
• Customer strategy
• Data & technology strategy
Da Data
• Engineering & architecture
• Management & governance
• Literacy & analytics
Ai Artificial intelligence
• Machine learning
• Generative AI
• Intelligent products
Cl Cloud
• Migration & modernization
• Optimization
• Infrastructure
Si Systems implementation
• Customer applications
• Business applications
• Integration
Ex Experience strategy & design
• Customer & employee experience
• Product strategy, design & innovation
• Advertising & marketing transformation
Dp Digital product building
• Product engineering
• Platform engineering
• Data engineering & machine learning
Pd Planning & delivery
• Product management
• Project & program management
• Lean portfolio management
Oc Organizational change & talent
• Change management
• Talent development & learning
• Culture transformation
Op Operations
• Mergers & acquisitions
• Operating model design
• Process improvement
Su Sustainability
• Strategy & impact tracking
• ESG digital transformation
• Value chain decarbonization
Ps Privacy & security
• Governance, risk & compliance
• Data privacy
• Security

Figure 1: Slalom Services

Slalom New England is a trusted partner in driving transformation across sectors. Whether modernizing healthcare systems, enabling secure cloud environments, or advancing equity through AI and data innovation, our teams deliver human-centered solutions that improve lives and strengthen communities. From federal and state agencies to Fortune 500s, we bring deep expertise and local commitment to every engagement.

- **Slalom New England |** We are driving innovation, equity, and transformation across public health, education, infrastructure, and technology. Our teams have modernized healthcare systems, launched AI Centers of Excellence, and delivered enterprise platforms for clients like Cigna and Insulet. We've built grants management tools for MassDOT, advised on security directives for the Executive Office of Public Safety and Security, and modernized cloud infrastructure for federal agencies.
- **Education |** We support higher education institutions, school districts, and education agencies with strategy, technology, and transformation services tailored to their mission. Our work spans secure cloud environments, AI enablement, and data modernization—empowering educators and administrators to drive student success and institutional resilience.
- **Health and Human Services |** We collaborate with human services agencies to transform the delivery of programs that support vulnerable populations and improve community well-being. Our teams have worked on modernizing eligibility platforms, enhancing case management systems, and reimagining digital enablement across state and local organizations—enabling agencies to deliver essential services more efficiently and equitably.
- **Public Health |** We empower public health organizations to harness the power of data and technology to advance community health initiatives and improve population outcomes. Our recent work includes supporting Arizona's statewide public health

data modernization—delivering cloud-based infrastructure, secure data integration, and advanced analytics to help agencies monitor, respond to, and prevent public health threats more effectively.

- **Healthcare |** We partner with healthcare providers, payers, and health tech organizations to deliver pragmatic, data-powered solutions that improve care experience, affordability, and outcomes. With 2,000+ healthcare-experienced consultants and engagements across 480+ clients, Slalom is recognized globally for advancing fiercely human healthcare through digital transformation, AI innovation, and human-centered design.
- **Robotics |** Slalom's Element Lab in New York City is an innovation hub with a focus on robotics and other emerging technologies. Slalom collaborated with Amazon Robotics to build and sustain the Global Robotics University (GRU) platform for training employees in robotics-enabled fulfillment centers. Hitachi America, Ericsson, and Slalom collaborated on a proof of concept to develop an automated quality inspection solution at Hitachi Rail's Maryland facility leveraging SPOT, a robotic dog from Boston Dynamics.

Our strategic partnerships.

Slalom's strategic alliances are a cornerstone of our ability to deliver scalable, future-ready solutions. As a premier partner to industry leaders like AWS, Snowflake, and Databricks, we bring deep technical expertise and co-innovation capabilities that accelerate cloud transformation, unlock data value, and enable advanced analytics and AI. These partnerships allow us to tailor solutions to your unique needs—leveraging trusted platforms to drive agility, security, and measurable impact.

slalom | aws

Slalom is a Premier Tier Partner in the AWS Partner Network (APN), with experience delivering cloud-native solutions across private and public institutions. Our partnership with AWS is built on a shared commitment to innovation, security, and mission-driven outcomes. Through our Strategic Collaboration Agreement (SCA), we co-invest in transformation via joint workshops, proof of concepts, and accelerators.

From strategic partnerships to constituent impact.

Through strategic alliances with over 700 technology providers—including AWS, Snowflake, and Databricks and other leading cloud, data, and AI companies—we deliver solutions that align with our client's goals. These partnerships enable us to:

- **Accelerate adoption** of secure, scalable, and compliant cloud services across private and public entities.
- **Influence product development** to meet the evolving needs of state agencies, municipalities, and public and private institutions.
- **Deliver best-in-class architectures** that our client's needs for availability, elasticity, and cost-effectiveness.

These partnerships enable us to remain platform agnostic and to think critically about the unique needs of each of our clients.

- 10x Partner of the Year (System Integrator – North America, Migration, GenAI).
- 2,900+ AWS projects delivered across technology, data and analytics, and strategy.
- 1,300+ AWS experienced consultants with 2,000+ AWS certifications.
- Recognized for excellence across Global, UK, and Ireland regions.

slalom | snowflake

Slalom is a Snowflake Elite Services Partner, with extensive experience delivering cloud-native data and AI solutions for public and private sector organizations. Our partnership is built on a shared commitment to accelerating data modernization, enabling secure, scalable analytics, and driving mission-driven outcomes. Through our strategic collaboration, we co-invest in transformation via joint workshops, proof of concepts, and accelerators, including GenAI and Machine Learning enablement programs.

- 7x Partner of the Year, including 2025 Global AI Partner of the Year.
- 1,800+ Snowflake projects delivered across data strategy, cloud migration, analytics, and business transformation.
- 650+ Snowflake-trained consultants with 100+ Snowflake certifications.

slalom | databricks

Slalom is a top-tier partner in the Databricks Partner Network, with deep expertise in delivering data intelligence, AI, and lakehouse solutions across industries. Our partnership is anchored in a shared focus on innovation, scalability, and client success. Through our partnership, we co-invest in transformation via joint solution development, accelerators, and enablement programs.

- 6x Databricks Partner of the Year, including 2025 SLED Partner of the Year, Canada Partner of the Year, and AMER Industries Business Unit Partner of the Year.
- 400+ Databricks engagements delivered across data engineering, strategy, governance, GenAI, and machine learning.
- 400+ technically trained consultants with 200+ Databricks certifications.

A key differentiator of Slalom is that we possess the breadth and depth of expertise across hyperscalers, platforms, and capabilities. We are positioned to design the big picture, implement the requirements, and adjust as needed, in collaboration with MassTech as we learn more about the realities of the use cases for DCC.

Our approach.

Our approach is rooted in human-centered design—we start by understanding the needs, motivations, and challenges of users and stakeholders. We combine this design approach

with deep technical expertise in cloud-native architecture, data engineering, analytics, and artificial intelligences. Our teams are skilled at navigating complex regulatory environments, implementing robust data governance frameworks, and designing solutions that are secure, scalable, and sustainable.

- **Human-Centered Design |** Our design thinking methodology focuses on the design of solutions that are intuitive, accessible, and tailored to diverse user needs, driving adoption and engagement across the DCC's target audiences.
- **Cloud-Native & Modular Architecture |** Our expertise in cloud-native, modular architectures enables rapid deployment, scalability, and future extensibility is critical to support the DCC's evolving needs.
- **Data Governance Leadership |** Slalom has implemented data governance frameworks for complex, multi-stakeholder environments, prioritizing data quality, lineage, and compliance with FAIR principles and ethical guidelines.
- **AI Solution Development and Implementation, including Synthetic Data Generation & Responsible AI |** We are recognized leaders in deploying AI systems including synthetic data generation solutions, leveraging open-source and commercial tools to achieve results that are validated, fair, maintain privacy, mitigate bias, and support Responsible AI innovation.
- **Accelerated Engineering |** Our Accelerated Engineering practice transforms the software development lifecycle by embedding AI from vision to scaled operations. Our approach delivers working software, strategy, and clarity—often within 24 hours. Clients benefit from faster time-to-market, reduced engineering costs, and enhanced reliability through solutions like AI-assisted incident management, synthetic data generation, and agentic orchestration. With expertise in AWS, Snowflake, and Databricks, we help organizations unlock AI maturity across individual productivity, role-based enhancement, and enterprise-wide automation.

Proposed Approach to Data Commons

We have developed a tailored approach to building the Data Commons that helps MassTech achieve its goals, respects constraints, and proactively manages the inherent risks involved in launching a first-of-its-kind platform.

Our plan involves 4 major phases. We propose launching the first version of a live system in 13 weeks and continuing to add features throughout a 53-week project timeline. Appendix – RFP Milestones includes a visual of how the project phases map to the milestones as defined

in the RFP.

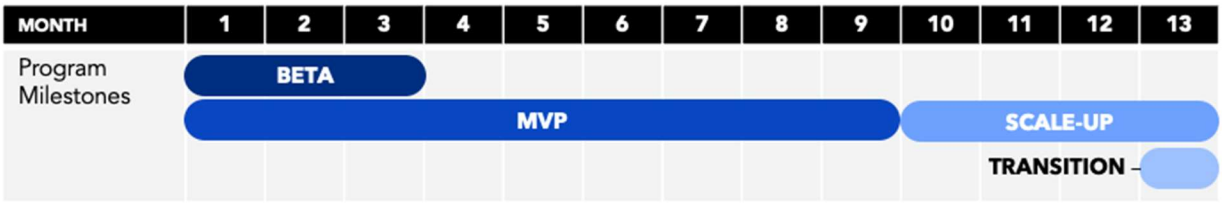


Figure 2: Project Timeline Overview

The below table outlines the objectives, summary, and features included in each of the 4 phases of the program.

PHASE	OBJECTIVE	SUMMARY	FEATURES
Beta	Rapidly learn about end-user behavior to understand adoption and usage patterns, tighten desired outcomes, and refine the roadmap.	Private, invite-only version of the Commons with limited functionality focused on access to data sets.	Data Catalog and Search, Simple Submission Form, Basic Governance Processes, Admin UI
Minimum Viable Product (MVP)	Build a critical mass of users and contributors with consistent meaningful use.	Production-ready public release of the Commons with leanest possible feature set required to achieve broad adoption, meaningful use and meet compliance requirements.	Basic Synthetic Data Tooling, Basic Fairness Tooling, Basic Credits System, Robust Security/Compliance, Enhanced Governance, Additional Data Sets and Data Modalities
Scale Up	Rapidly drive adoption, usage and referral by aggressively expanding data catalog, tool set and user base.	Frequent releases of new platform features, each tied to a specific value hypothesis (Adoption, Usage, Ethics, or Risk Mitigation).	Enhanced Synthetic Data Tools, Enhanced Fairness Tools, Enhanced Credits System, Additional Data Sets and Advanced Data Modalities
Transition to Steady State	Empower Data Commons team to run the platform and supporting operations to sustainably run a program that delivers value to the Commonwealth for the long term.	Transition technical ownership from Slalom Project Team and enable a sustainable Operating Model for the Commons (service, governance, admin, finance/accounting, etc.)	Continued feature and data set enhancement

This balanced approach prioritizes a working product, in the hands of users, within the first 3 months and to meaningful scale in the first year. A phased release approach accelerates learning and drives frictionless adoption, while remaining cost-conscious, building only what is needed. This approach enables us to deliver value early, reduce the risk of overbuilding, and allows us to adapt based on real user feedback—making the product practical and sustainable.

Note on Scope and Cost Flexibility

Our current estimate for delivering all phases and features described is approximately **$15 million** (please see Attachment C – Services Budget Template for a detailed cost breakdown), reflecting a comprehensive build with integrations across all major components. However, the DCC's scope is inherently flexible. Our proposed architecture and phasing strategy allow for a wide range of investment levels without compromising quality or user experience.

If desired, the solution can be right-sized during project execution through measures such as hosting a smaller number of curated datasets, linking to external sources rather than fully hosting them, implementing a simplified Commons Credits mechanism, or focusing on high-value domains and modalities first. These options will be explored collaboratively during beta to align scope, budget, and impact. Our phased approach is intentionally designed to preserve agility—enabling MassTech to scale up or down based on priorities, outcomes, end user learnings, and available funding.

These core principles guide our solution:

- **Outcome-Oriented:** The Data Commons is rooted in the exciting objective of empowering researchers and innovators; keeping Massachusetts at the forefront of innovation, research, founding and growing businesses. We work backwards from the stated goal to "what must we do to achieve that outcome?" In our experience, the alternative approach—to make a list of requirements and outline a plan to check them off—is susceptible to budget overruns and poor execution.

- **Human-Centered:** We specialize in the technologies described in the RFP, and we firmly believe that best-in-class execution of a technology build is *necessary but not sufficient* to achieve the goals of the Data Commons. The technology will only be valuable if it meets a real need for end users and is usable. We created our technical solution by examining end user needs, rapidly mocking up an interactive user experience, and then identifying the technical components needed to make it real.

Flexible, Outcome-Driven, and Cost-Conscious Delivery.
We estimate a **$15 million** investment for a full-featured build, but our agile architecture allows for scalable delivery without compromising quality. The solution can be right-sized during execution to align with budget and impact—by hosting fewer datasets, linking to external sources, or prioritizing high-value domains. Guided by outcome-oriented, human-centered, and risk-aware principles, our approach prioritizes clarity, usability, and adaptability throughout the project.

- **Risk-Aware:** Executing an innovative program is fundamentally different from building technology that has many precedents to draw on. We must accept that there is intrinsic uncertainty about exactly how users will behave once the Data Commons exists. We account for that uncertainty by using the tried-and-true method of identifying key challenges and *pulling them forward* in the project plan, so we confront them as soon as possible. That way, if there are major assumptions in our plan that prove to be false, we have the resources to adjust and find alternate routes to achieve our goals. Slalom's Accelerated Research Centers (ARC) specializes in de-risking innovation concepts through practical, real-world experimentation that precedes even the capital investment of alpha and beta phases. With a staff of game theorists, behavioral psychologists, and designers, ARC uses a combination of prototyping and stagecraft to prove and disprove hypotheses that are key to market-making concepts.
- **Cost-conscious:** This program could be very large or very lean, depending on the depth and breadth of features. Complex programs run into major cost problems when there is poor execution or overbuilding. In our experience, these problems are rarely driven by technical challenges. They are more often driven by a lack of clarity and alignment, a lack of focus on outcomes, and failure to rigorously consider end user behavior. We believe the most important lever for controlling cost is controlling scope. We recommend proactively and relentlessly driving clarity for all stakeholders around what is being built, for whom, and why; embracing a "less is more" mentality. This discipline benefits your team, our team, and our shared end users. We take our commitment to your success seriously and expect to productively challenge ourselves and you to build only what is needed to achieve the outcomes you aspire to.

Phasing Strategy

The Data Commons is an innovative, first-of-its-kind platform. With innovation comes uncertainty and risk. The Data Commons faces both technical and strategic risks; some – like data security, privacy, and compliance – are critical but well understood, with mature safeguards such as encryption, access tiering, and multi-factor authentication built into the solution from the start. Others are less predictable because they rely on future behavior: adoption, sustained engagement, and long-term sustainability. These "behavioral" risks are often the real differentiators between projects that succeed and those that stall. Our phased approach focuses on retiring those strategic risks early, while managing known technical risks through established best practices.

The concept of **phased risk retirement** is a disciplined way to address these challenges *without limiting ambition*.

Rather than trying to build everything at once, we sequence development around the **behavioral risks that matter most**. Each phase is designed to build confidence and intentionality in decisions made to retire a specific set of risks before committing additional resources. By addressing risks in order, and as inexpensively as possible, we conserve scarce resources for later phases, when scale and robustness will matter most and uncertainty about end user behavior has been substantially reduced.

This approach reduces wasted investment, accelerates learning, and when we do scale, allows us to scale something differentiated and defensible that people demonstrably want.

The Biggest Risks (in Order)

RISK	SUMMARY	MITIGATION STRATEGY	TIMELINE
Adoption and value proposition	Platform may not attract users or offer a value proposition that differentiates it from existing alternatives.	Low-cost, scientifically sound experiments followed by a private beta early in the project.	Addressed during the Beta phase.
Critical Mass	Foster sufficient adoption, frequent users, and strong referral behavior to maintain an active and engaged user base.	Leverage network effects– make early users successful and visible, provide reasons to return and bring others. Create direct user engagement to understand motivations and hurdles.	Addressed during the MVP phase.

RISK	SUMMARY	MITIGATION STRATEGY	TIMELINE
Overconsumption	Prevent over-consumption and under-contribution to support the Commons' ongoing growth in both quality and participation.	Phased and/or tiered approach to the Credits system.	Addressed during the MVP phase.
Over-engineered solution	Minimize building more than what is required by the users.	Prioritize each new release around a clear, testable assumption regarding what users will value most, and expand after that value is proven with real use.	Addressed during the Scale Up Phase.
Impact	Realize original vision by leveraging adoption and real-world user data.	Measure whether the platform is advancing the Commonwealth's goals as desired, and adjust roadmap or policies based on data.	Addressed during the Scale Up Phase.
Continuity	For the platform to deliver on its mission for the long-term, there must be an organization with clear roles, processes, and governance rules in place to continue to maintain the platform technology, administer the platform, serve users, and drive continued adoption.	Work collaboratively with the MassTech team during live operations, observe how the platform runs in practice, and use insights to design the future-state organization, roles, and governance model.	Addressed in the Transition phase.

Figure 3, below, depicts the anticipated risks and the timeline in which we expect for them to be addressed.

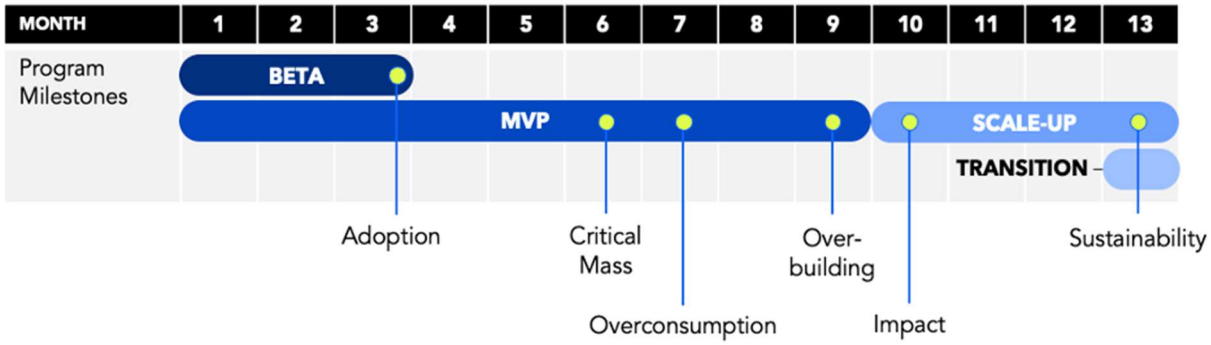


Figure 3: Project Timeline With Risks

Solution Summary

Overview

The Massachusetts AI Hub Data Commons connects researchers, agencies, startups, and civic groups to trustworthy, AI-ready data in one place. A unified portal, open catalog, and modular services—such as fairness checks, synthetic data generation, and Commons Credits—turn safe access into everyday use.

This solution is designed to serve **thousands of engaged users** with diverse missions and technical backgrounds while remaining **economically sustainable for the Commonwealth**. Built with a **phasing strategy**, the Commons grows in capability over time, enabling early value delivery while avoiding overbuild and supporting long-term scalability.

Concept of Operations

The Concept of Operations illustrates how the Data Commons functions as a seamless whole—connecting front-end user interactions to the layered system architecture behind them. Each step in the user journey (from login to data submission) draws on multiple architectural layers—Experience & Access, Control, Data & Compute, Curation, and Admin—working in concert. While users experience a single, intuitive interface, every function they touch activates a coordinated set of back-end services that support security, governance, and performance at scale. A more detailed depiction of the underlying system architecture can be found in **Figure 5** of this proposal.

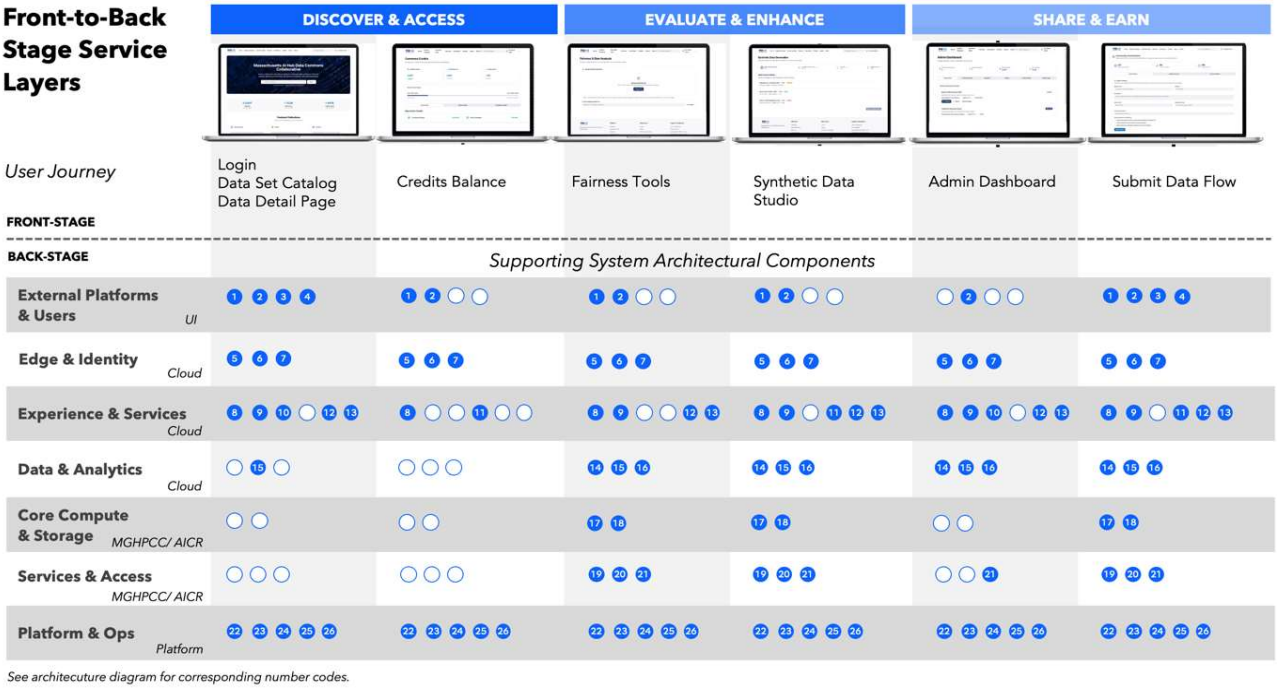


Figure 4: User Interactions and System Architecture

Technology Architecture Design

The Technical Architecture diagram provides a system-level view of how the Data Commons is constructed. It focuses on the architecture itself—showing how modular, API-connected components exchange data securely across layers. Together, they enable scalability, interoperability, and the flexibility to evolve without replatforming.

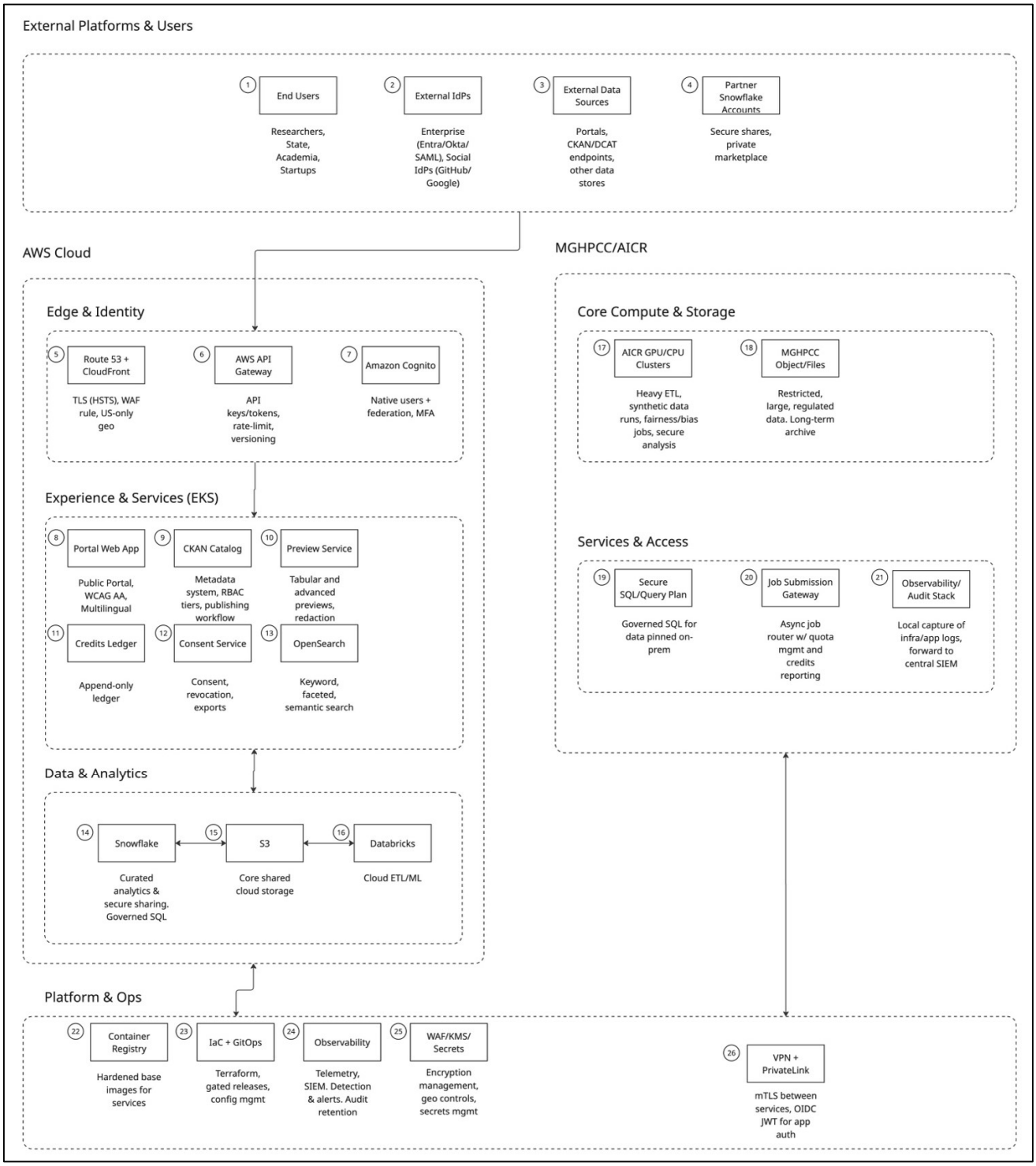

External Platforms & Users
1 End Users
Researchers, State, Academia, Startups
2 External IdPs
Enterprise (Entra/Okta/ SAML), Social IdPs (GitHub/ Google)
3 External Data Sources
Portals, CKAN/DCAT endpoints, other data stores
4 Partner Snowflake Accounts
Secure shares, private marketplace
AWS Cloud
Edge & Identity
5 Route 53 + CloudFront
TLS (HSTS), WAF rule, US-only geo
6 AWS API Gateway
API keys/tokens, rate-limit, versioning
7 Amazon Cognito
Native users + federation, MFA
Experience & Services (EKS)
8 Portal Web App
Public Portal, WCAG AA, Multilingual
9 CKAN Catalog
Metadata system, RBAC tiers, publishing workflow
10 Preview Service
Tabular and advanced previews, redaction
11 Credits Ledger
Append-only ledger
12 Consent Service
Consent, revocation, exports
13 OpenSearch
Keyword, faceted, semantic search
Data & Analytics
14 Snowflake
Curated analytics & secure sharing. Governed SQL
15 S3
Core shared cloud storage
16 Databricks
Cloud ETL/ML
MGHPCC/AICR
Core Compute & Storage
17 AICR GPU/CPU Clusters
Heavy ETL, synthetic data runs, fairness/bias jobs, secure analysis
18 MGHPCC Object/Files
Restricted, large, regulated data. Long-term archive
Services & Access
19 Secure SQL/Query Plan
Governed SQL for data pinned on-prem
20 Job Submission Gateway
Async job router w/ quota mgmt and credits reporting
21 Observability/ Audit Stack
Local capture of infra/app logs, forward to central SIEM
Platform & Ops
22 Container Registry
Hardened base images for services
23 IaC + GitOps
Terraform, gated releases, config mgmt
24 Observability
Telemetry, SIEM. Detection & alerts. Audit retention
25 WAF/KMS/ Secrets
Encryption management, geo controls, secrets mgmt
26 VPN + PrivateLink
mTLS between services, OIDC JWT for app auth

Figure 5: Technical Architecture

Example User Journey

Together, this interplay between user experience and system architecture makes complex capabilities—from dataset discovery to fairness auditing and synthetic generation—feel unified and intuitive to the end user. The following example brings that experience to life through a realistic user journey.

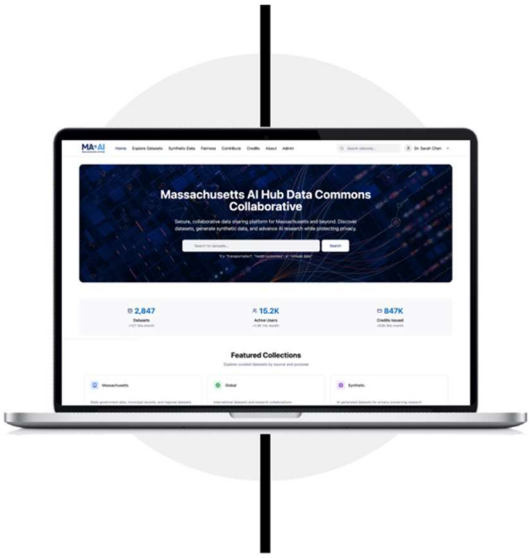
Envisioning the Solution

The following vignette illustrates how a Massachusetts user might engage with the Data Commons to solve a real-world problem. It highlights representative features and workflows, not an exhaustive list. The user interface shown here is conceptual; final design will be developed collaboratively during the project. The scenario reflects functionality expected in the mid-to-late Scale-Up phase of the program.

Meet Marcus, a Senior Machine Learning (ML) Engineer at a Boston-based autonomous vehicle startup. His team is developing perception algorithms for urban driving scenarios with his team, but they've discovered their current training dataset severely underrepresents challenging conditions like night driving and adverse weather. With safety-critical applications, Marcus needs to confirm that their models perform equitably across all driving conditions before deployment.

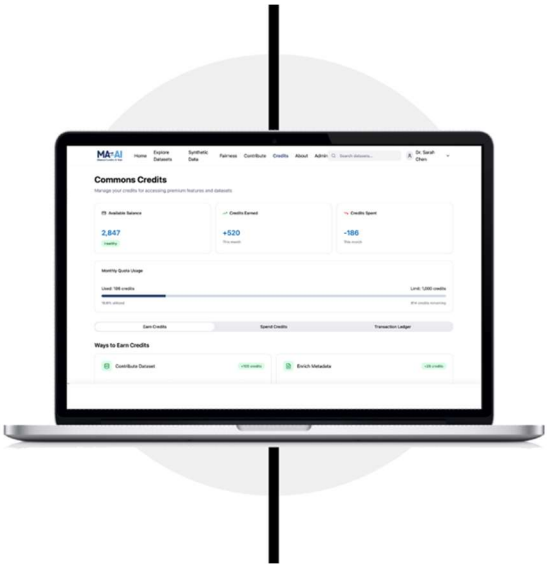
Discovery & Access

Marcus visits the Massachusetts AI Hub Data Commons and logs into his account.



Credits Dashboard

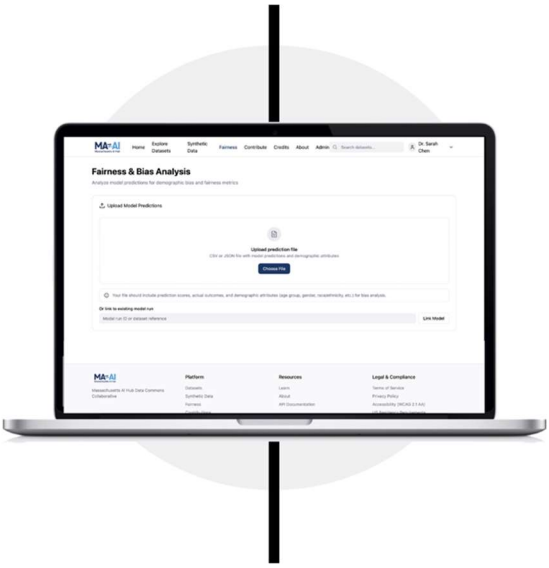
His Credits dashboard shows his team's Commons Credits balance and recent activity.



Fairness & Bias Analysis

He navigates to upload his autonomous driving dataset — 50,000-point cloud frames from Boston-area test drives containing LiDAR scans, camera feeds, and labeled objects. After upload, Marcus receives a notification that the AI Fairness and Bias Capability has completed its audit. The fairness dashboard reveals critical gaps:

- Only 10% of frames represent night conditions (vs. 40% real-world driving)
- Just 5% include rainy weather (vs. 15% regional average)
- Severe underrepresentation of pedestrians in low-visibility conditions

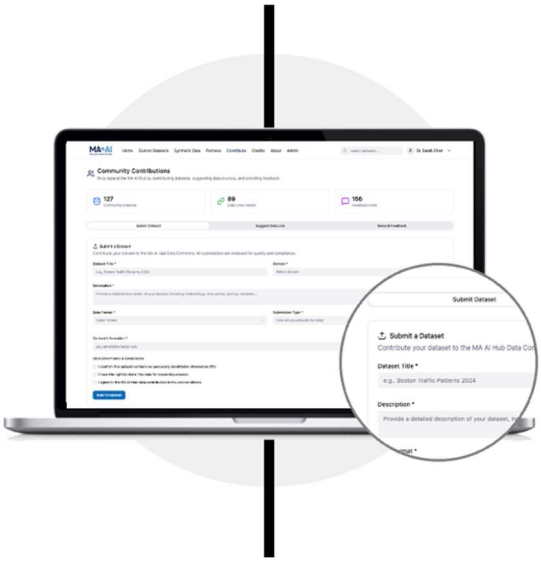


Synthetic Data Generation

Marcus uses the CARLA simulator to generate synthetic data targeting these gaps. The system provides real-time progress tracking as quotas are filled: night conditions increase from 10% to 25%, rainy weather from 5% to 15%. The fairness tool continuously re-audits, and once quotas are met, automatically generates a Fairness Card documenting the complete bias mitigation process.

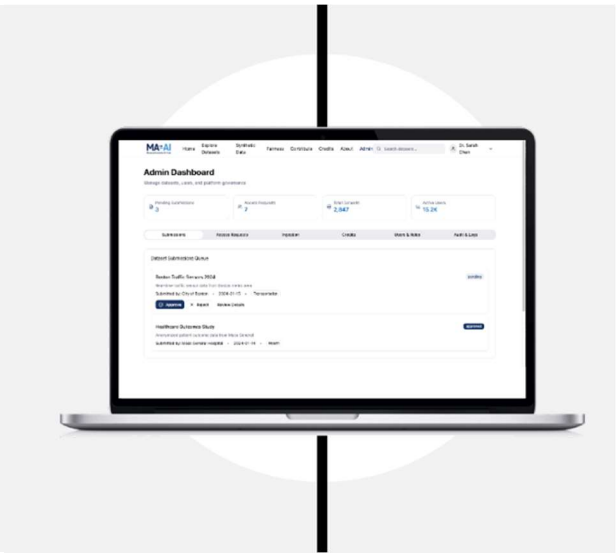
Sharing & Earning Credits

With his enhanced dataset now containing 65,000 properly representative frames, Marcus contributes back to the ecosystem. He uploads the improved dataset with full fairness documentation, shares his CARLA configuration templates for other Boston-area AV teams, and publishes a case study demonstrating fairness-driven development practices. His dataset enters the administrative queue for review and approval.



Administrative Review

The Data Governance Administrator receives a notification that Marcus's enhanced dataset has been submitted for community sharing. From the Admin Dashboard, they review the submission, confirm it meets DCC standards for community sharing, and approve it for public availability. They add it to the featured datasets with a "Community Validated - Fairness Enhanced" badge and award Marcus additional Commons Credits for his valuable contribution.



This intuitive experience has allowed Marcus to discover bias automatically, generate synthetic data to fill gaps, and validate fairness improvements, transforming weeks of work into a seamless, integrated process. This journey illustrates how the DCC integrates fairness, synthetic data, and community contribution into one cohesive workflow—turning responsible AI from aspiration into everyday practice.

"We've not only solved our bias problem, but we're helping the entire Massachusetts AV community build safer, more equitable systems."

Future State Vision: From Portal to Intelligent Network

While the proposed solution delivers a world-class Data Commons today, its modular, API-first architecture also lays the groundwork for tomorrow's innovations. Over time, the DCC could evolve from a portal into a network of intelligent, trustworthy agents that act on behalf of each participant (e.g., researchers, agencies, startups, and the public). In this future vision, people wouldn't navigate menus or forms; instead, they would express intents ("make this dataset available responsibly", "analyze pedestrian safety near schools") and the system's agentic layer would handle the orchestration.

Behind the scenes, these agents would coordinate dataset discovery, permissions, compliance, compute, and collaboration—surfacing options, costs, and risks transparently for human approval. Slalom has already delivered similar experiences in other industries, such as a public-facing conversational AI chatbot for Norwegian Cruise Lines that uses Agentic AI to turn a user's travel intent that they express in natural language into a fully planned itinerary through secure, explainable automation with a human in the loop. This same paradigm could one day enable DCC users to describe a research or policy goal in plain natural language and watch as an intelligent agent securely discovers, assembles, and validates the required data, automating weeks of technical orchestration into minutes of review.

While this agentic evolution is **not included in the initial scope**, the solution is intentionally architected so that the Commonwealth could pursue this path, should it choose to, without

needing to replace the fundamental architecture that is in scope of this project. The same microservices foundation that will deliver a working Beta in 13 weeks can, over time, evolve seamlessly into this intelligent ecosystem. **No replatforming or major refactoring is required, only continuous refinement and expansion of capability.**

Team Structure

We have structured the team around clear accountability and cross-functional collaboration. Each group owns a distinct **dimension of value creation**—from defining direction and user needs, to building the platform and enabling adoption—while staying tightly aligned through an integrated program office. This design prioritizes clarity, speed, and a seamless experience for both MassTech and end users.

Our team structure is built on a foundation of deep integration with MassTech leadership. By embedding regular touchpoints, shared decision-making forums, and transparent communication channels, we position our teams to remain aligned with MassTechs' vision and evolving priorities. This close partnership enables us to proactively identify and manage risks together, adapt to new information, and make timely course corrections as the program unfolds.

Strategy & Program Team (~4 FTE)
This team sets the direction, manages scope, and keeps all workstreams aligned to outcomes. It focuses on de-risking, tracking assumptions, and keeping success measures front and center. The team includes an Accountable Executive, a Chief Product Owner, an Engagement Lead, and a Senior Solution Architect. Together, they serve as the connective tissue across design, engineering, and adoption, providing a coherent program and a consistent client experience.

Integrated, Outcome-Driven Team Structure
Our delivery model is built on clear accountability and cross-functional collaboration, with specialized teams aligned to distinct dimensions of value creation. Four core groups — Strategy & Program, Design, Technology, and Adoption & Enablement — work in concert to define direction, build the platform, and drive adoption. This structure enforces clarity, speed, and a seamless experience for MassTech and end users, while enabling agility and alignment throughout the program.

Design Team (~2 FTE)
This team is responsible for understanding end-user needs and translating them into intuitive, testable experiences. It includes a Design Strategist, who leads user research, early prototyping, and information architecture, and a UX/UI Designer, who shapes the detailed interface design. Their work

grounds the technology, building in real user behavior, and makes the product as usable as it is powerful.

Technology Teams (~40 FTE)
The technology build is delivered by four full-stack engineering pods—focused on Product & API Experience, Data & Discovery, Platform & Trust, and Shared Services. This structure maximizes throughput, quality, and flexibility while maintaining cost realism through a balanced onshore/nearshore delivery model. Select nearshore engineers—operating in East Coast time zones—provide extended coverage and efficiency without compromising collaboration or velocity.

All resources adhere to the Commonwealth's guidelines for domestic versus foreign participation in government projects, supporting compliance while optimizing value. Each pod includes product, engineering, and QA capabilities, supported by cross-workstream leads in Quality Engineering, Testing, and Site Reliability. Specialized Cloud, Data, and Solution Engineering SMEs join as needed to address niche technical challenges efficiently. The number of team members, pods, and pod composition could vary based on the final scope and roadmap as the project evolves.

Adoption & Enablement Team (~3 FTE)
This team fosters engagement with the platform, both externally and internally, aiming for more than just delivery to users. Externally, it drives adoption through end-user training, communications, and engagement. Internally, it designs the future-state operating model, governance, and processes needed for long-term sustainability. Roles include an Adoption & Enablement Lead, an Operating Model Lead, and a part-time Subject Matter Expert for specialized support, with a Training Architect embedded in the engineering pod to support smooth handoffs between build and rollout.

Together, these teams form a cohesive, outcome-driven organization that moves in unison—building, learning, and adapting alongside MassTech to deliver a platform that is as enduring as it is impactful.

Please reference Appendix – Proposed Team Structure for a more readable version of the team structure.

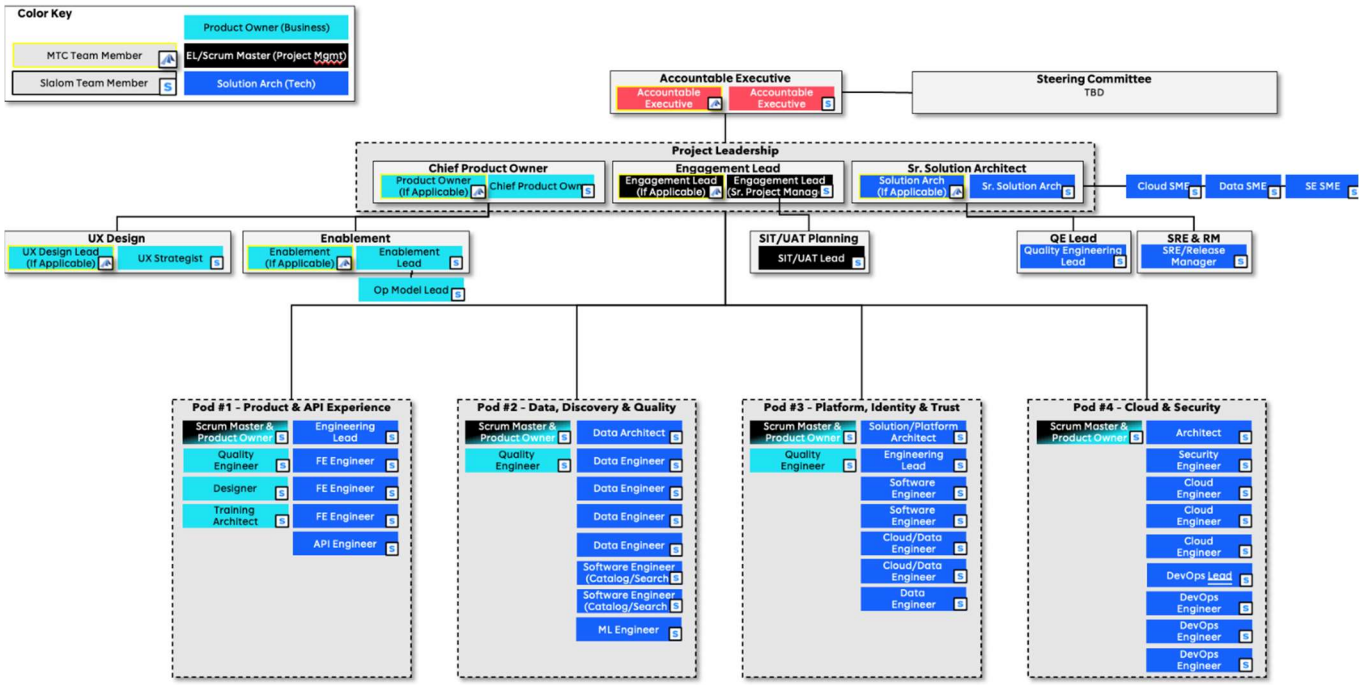


Figure 6: Proposed Full Team Structure

Collaboration & Delivery Model

Our Collaboration Philosophy: We Work With You, Not At You

At Slalom, we believe that the most successful projects are built on a foundation of true partnership. Our approach is rooted in the simple but powerful idea that we are here to work *with* you, not *at* you. We integrate into your teams, fostering a collaborative environment where we share our knowledge and empower your people to own the solution long after we're gone.

Here's what that means in practice:

- **One Team, Shared Goals:** We embed ourselves within your organization, working together shoulder-to-shoulder. We share a common set of objectives and work together to achieve them. Your success is our success.
- **Transparency and Open Communication:** We prioritize honest and open dialogue. You will continuously have a clear view of our progress, challenges, and next steps. We believe in surfacing issues early and tackling them collaboratively, so that there are no surprises along the way.
- **Workshop Facilitation and Co-Design:** We don't just present findings; we facilitate consensus where possible and clear decisions when they're needed. We use interactive techniques to bring diverse viewpoints together and drive toward shared understanding. Through expert workshop facilitation and a "show-don't-tell" mindset,

we make abstract concepts tangible. Our interactive "co-design" sessions make all stakeholders active participants in shaping the solution, which builds buy-in and helps the outcome reflect the collective needs of the group.

- **Building Your Capabilities, Not Ours:** Our goal is to make you self-sufficient. We are committed to knowledge transfer and mentorship so that your team is equipped with the skills and understanding needed to carry the work forward.
- **Flexibility:** We understand that business needs can change. Our methodology is agile and flexible, allowing us to adapt to new information and shifting priorities.

By working in this deeply collaborative way, we deliver more than just a project; we deliver lasting value and a stronger, more capable organization.

Product Engineering Methodology

Slalom's Product Engineering Methodology (PEM) is a structured, phased framework that integrates modern engineering practices, agile delivery, and stakeholder collaboration to drive end-to-end product development. This methodology provides a consistent, tested approach for delivering software incrementally with quality and speed. It has the flexibility to integrate with MassTech's processes and uses a business value-focused approach so that MassTech is prepared to successfully operate the DCC for the long-term.

Collaborative, Transparent, and Empowering Delivery Model
Slalom's delivery model is engineered for transparency and adaptability—ideal for the development of the DCC. Our proprietary Product Engineering Methodology (PEM) integrates agile delivery with modern engineering practices to accelerate outcomes while maintaining technical rigor.

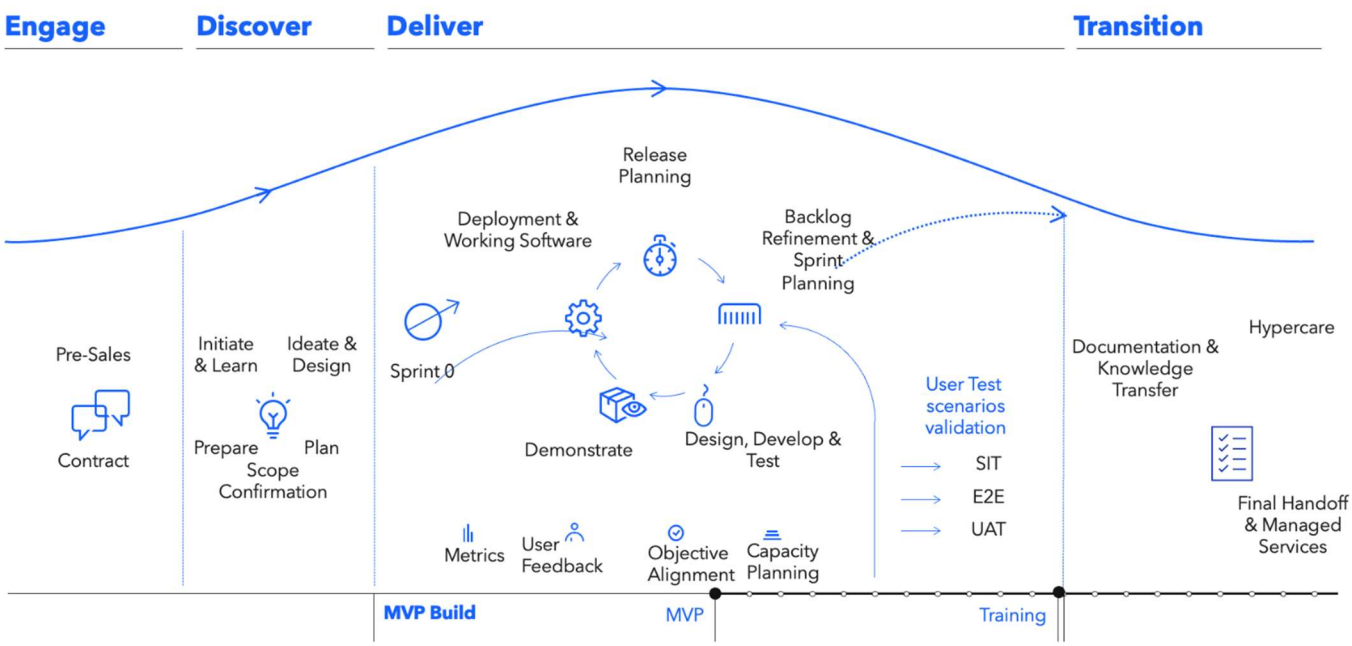


Figure 7: Product Engineering Methodology Process

Our approach relies on multi-functional agile teams, including engineering, quality engineering, DevOps, and user experience, to provide a high-quality product with reliability, quality, and security. **The intent of PEM is to create Minimum Viable Products (MVP) in an agile fashion to provide value to the users of the DCC as quickly as possible.**

At its core, PEM begins with Engage, where we align with MassTech early to shape expectations and establish a shared vision. This phase sets the tone for collaboration and positions leadership teams to be embedded from day one. The discover phase focuses on validating product feasibility and viability. For the DCC, we will combine this user discovery with a beta release so that we can capture feedback of real-world users. The deliver phase transforms feedback and plans into a more robust DCC MVP. We leverage agile principles and will partner with MassTech to release to users with speed and quality. The final transition phase focuses on a smooth handoff to MassTech, moving ongoing development and support into a sustainable model.

PEM is supported by a rich library of templates and artifacts—from journey maps and persona templates to ecosystem diagrams—that guide teams through each phase. It is a methodology that not only delivers products but also build capabilities, enabling our clients to operate and evolve their solutions with confidence.

Our Commitment to Delivery: “The Big 3 for Delivery Success”

We understand that even well-planned projects can encounter challenges. The reasons are often predictable and manageable: unclear expectations, creeping scope, poor communication, or a lack of oversight. At Slalom, we have a proactive and disciplined approach to project delivery designed to mitigate these aspects from day one. We call it our **"Big 3 for Delivery Success."**

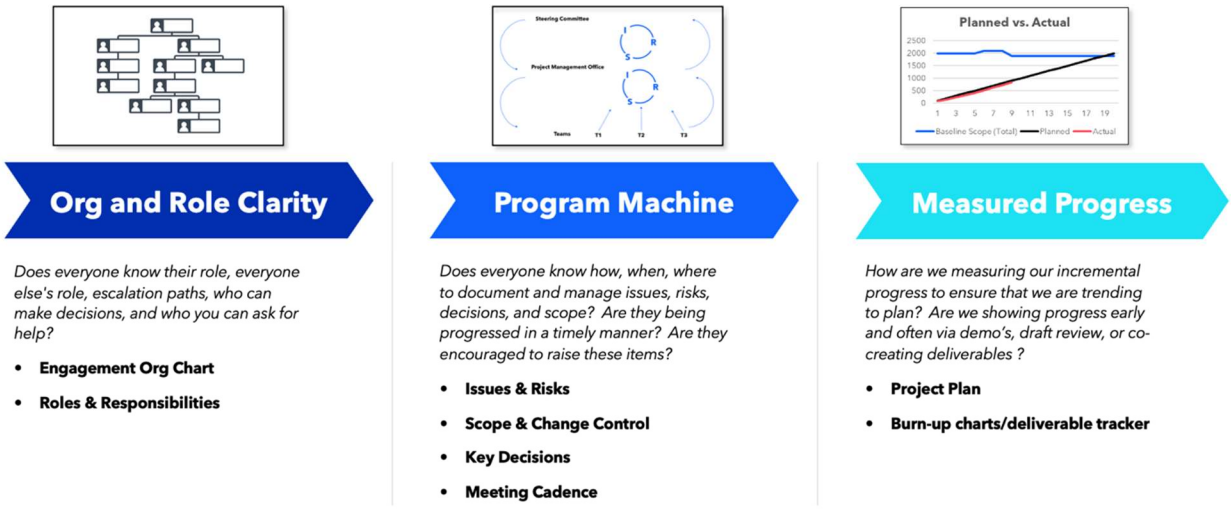

Org and Role Clarity
Does everyone know their role, everyone else's role, escalation paths, who can make decisions, and who you can ask for help?
Engagement Org Chart
Roles & Responsibilities
Program Machine
Does everyone know how, when, where to document and manage issues, risks, decisions, and scope? Are they being progressed in a timely manner? Are they encouraged to raise these items?
Issues & Risks
Scope & Change Control
Key Decisions
Meeting Cadence
Measured Progress
How are we measuring our incremental progress to ensure that we are trending to plan? Are we showing progress early and often via demo's, draft review, or co-creating deliverables ?
Project Plan
Burn-up charts/deliverable tracker
Planned vs. Actual

Figure 8: The "Big 3" for Delivery Success

This framework is built on three core pillars that keep our projects on track, on budget, and aligned with your goals.

1. **Organizational & Role Clarity |** To prevent confusion and support clear lines of communication, we begin every engagement by establishing a clear organizational structure. Every person on the project team, from both Slalom and your organization, will understand their specific role, the roles of others, and the established escalation paths for resolving issues. We define who holds decision-making authority and who to turn to for help, creating a transparent and efficient team environment from the start.
2. **The Program Machine |** Successful projects run on a well-defined operational rhythm. We establish a "Program Machine" that provides a clear and consistent framework for managing the project's core mechanics. These include systems for managing Issues & Risks, Scope & Change Control, Key Decisions, and Meeting Cadence.
3. **Measured Progress** | Progress must be visible, tangible, and understandable. To confirm that we are consistently trending toward our goals, we measure our progress incrementally and share it with you frequently. This is achieved through **Modular Project Plans, Visual Burn-up Charts & Deliverable Trackers, and Frequent Demonstrations of working software.**

Program Governance & Stakeholder Management

A successful project requires more than just a strong delivery plan; it demands a governance structure that keeps all parties aligned, informed, and engaged. Our governance model is built on a foundation of proactive communication and transparent oversight, designed to keep the program on track and manage stakeholder expectations effectively.

- **Structured stakeholder engagement |** We begin by mapping out all key stakeholders to understand their roles, interests, and communication needs. This

allows us to create a tailored engagement plan so that everyone from the project team to executive leadership receives the right information at the right time.

- **Project kick-off |** We will kick off the project with leadership and key stakeholders to review and align program objectives, guiding principles, team organization, and communication processes so the team can start strong and work through critical decisions early.
- **Steering committee for strategic oversight |** We will establish a Steering Committee comprised of key leaders from both your organization and Slalom. This group will meet on a regular cadence to provide strategic guidance, make key decisions, remove roadblocks, and keep the program aligned with your objectives.
- **Detailed project plan |** Our project manager will be responsible for defining and holding the team accountable to a plan that outlines the key milestones, activities, and dependencies, with assigned owners.

Status Reporting

Slalom's project manager will lead a weekly status report meeting with MassTech and Slalom leadership. During this meeting, we will discuss progress against our plan, risks and issues, and key decisions that need to be made. The figure below depicts what a typical status report looks like. In addition to the below status report, we will provide a monthly financial report, so that MassTech has clarity on planned versus actual spend by role and hours.

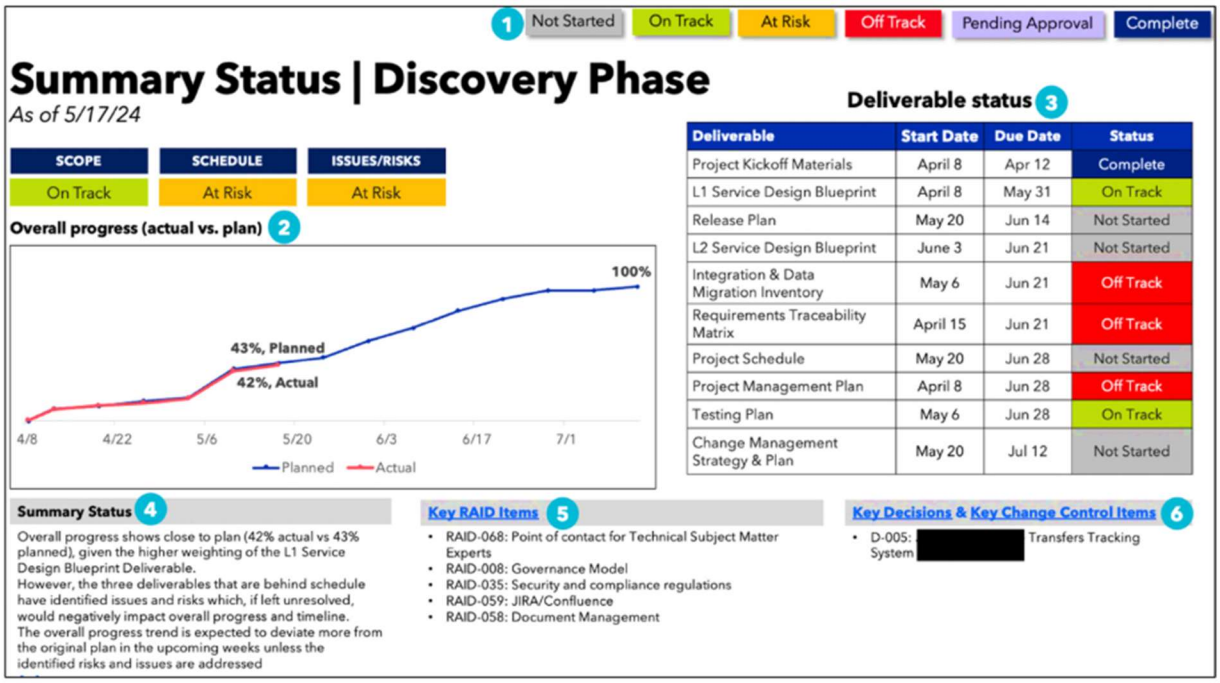

1
Not Started
On Track
At Risk
Off Track
Pending Approval
Complete
Summary Status | Discovery Phase
As of 5/17/24
Deliverable status 3
SCOPE
SCHEDULE
ISSUES/RISKS
On Track
At Risk
At Risk
Overall progress (actual vs. plan) 2
100%
43%, Planned
42%, Actual
4/8
4/22
5/6
5/20
6/3
6/17
7/1
Planned
Actual
Deliverable | Start Date | Due Date | Status
Project Kickoff Materials | April 8 | Apr 12 | Complete
L1 Service Design Blueprint | April 8 | May 31 | On Track
Release Plan | May 20 | Jun 14 | Not Started
L2 Service Design Blueprint | June 3 | Jun 21 | Not Started
Integration & Data Migration Inventory | May 6 | Jun 21 | Off Track
Requirements Traceability Matrix | April 15 | Jun 21 | Off Track
Project Schedule | May 20 | Jun 28 | Not Started
Project Management Plan | April 8 | Jun 28 | Off Track
Testing Plan | May 6 | Jun 28 | On Track
Change Management Strategy & Plan | May 20 | Jul 12 | Not Started
Summary Status 4
Overall progress shows close to plan (42% actual vs 43% planned), given the higher weighting of the L1 Service Design Blueprint Deliverable.
However, the three deliverables that are behind schedule have identified issues and risks which, if left unresolved, would negatively impact overall progress and timeline.
The overall progress trend is expected to deviate more from the original plan in the upcoming weeks unless the identified risks and issues are addressed
Key RAID Items 5
RAID-068: Point of contact for Technical Subject Matter Experts
RAID-008: Governance Model
RAID-035: Security and compliance regulations
RAID-059: JIRA/Confluence
RAID-058: Document Management
Key Decisions & Key Change Control Items 6
D-005: [redacted] Transfers Tracking System

Figure 9: Sample Status Report

The status report consists of:

1. **Status indicators** serve as a visual cue about the progress or status of deliverables and development, allowing for quick comprehension and decision making.
2. **Measured progress** allows us to track our actual progress against our planned progress and anticipate/evaluate potential impacts to scope or schedule. The method of measurement will be based on the discovery plan and schedule that will be defined in detail within the Initiate & Plan phase of the project.
3. **Milestone status** represents a quick view into the progress of each milestone.
4. **Summary status** provides an executive-level summary of overall progress
5. **Key RAID Items** summarize RAID (risks, assumptions, issues, and dependencies) that are impacting the overall progress of the project.
6. **Key Decisions & Change Control** provides a glimpse into items that require discussion to reach a conclusion regarding scope.

These structures, when combined with structured communication cadence and "Show-Don't-Tell" demonstrations, prevent surprises. It provides clear channels for communication, forums for leadership to make key decisions, and a consistent flow of information that keeps all parties confidently aligned from kickoff to completion.

Design Approach

Our design approach is how we make the Data Commons real. We focus the team's effort where it matters most to end users: creating a platform people will use and love. Adoption (signing up for the platform) and Usage (engaging with it meaningfully over time), Contribution (sharing data and tools), and Referral (bringing others onto the platform) are the end user behaviors that will ultimately drive the Commonwealth towards its goals.

Our design teams use cutting-edge AI-assisted methodologies to accelerate without losing focus on the core practices of human-centered design that have proven critical to modern technology development.

The below image depicts how the design elements are layered into our overall timeline.

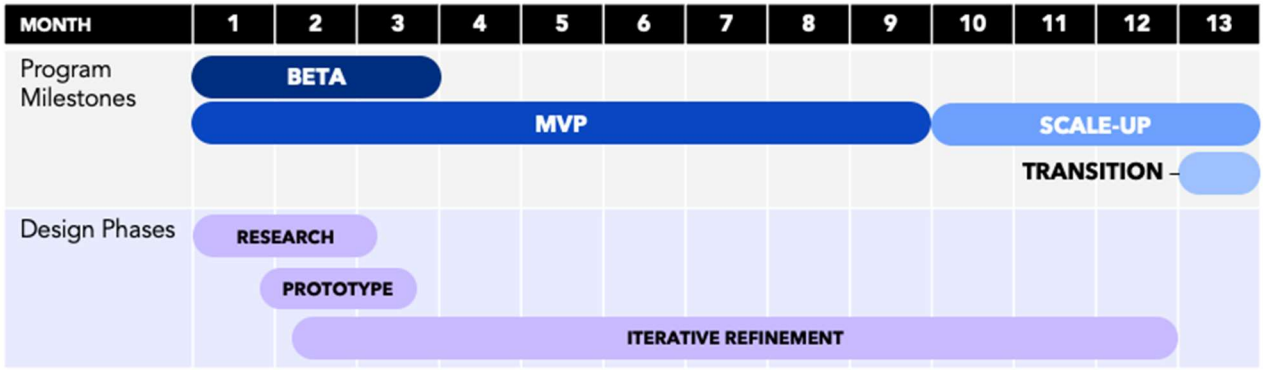


Figure 10: Project Timeline With Design Phases

Rapid Discovery (Beta Phase)

In the first 6 weeks of the program, we will conduct comprehensive user research to establish a clear, shared understanding of who we're building for and what they need most. This foundational work is critical to building the right features in the right sequence, and assuring that our team and your team remain aligned throughout delivery.

Human-Centered, Rapid, and Iterative Design Approach.
Our design approach puts end users at the center, resulting in a Data Commons that is intuitive, engaging, and widely adopted. We start with comprehensive user research to define personas and needs, then leverage AI-assisted prototyping to quickly align stakeholders on the user experience before development begins. Throughout the project, design is embedded in sprint teams, operating ahead of development to refine features based on real user feedback and usage analytics. Every release is discoverable, usable, and impactful—driving adoption, contribution, and referral across a diverse innovation community.

Personas

The diversity of the anticipated Data Commons user base (e.g., AI/ML developers, academic institutions, state government agencies, start-ups and enterprises, and civic tech groups) makes for a thriving innovation community; and requires careful consideration of where each group has unique needs and where their needs overlap.

We will codify the findings from our interviews into **persona profiles** representing the full spectrum from AI researchers to civic organizations, prioritizing them based on impact potential and platform adoption. When we deeply understand our end users and their needs, we reduce ambiguity and grow confidence, resulting in better software and well-controlled scope.

Accelerated Prototyping

We will create a set of **interactive prototypes** that demonstrate the conceived user journeys across the three development phases (Beta, MVP and Scale-Up) before significant resources begin development. Slalom's proprietary AI-Assisted Design tools will allow us to generate all 3 versions of the prototype in days, and our expertise in workshop facilitation will help align all stakeholders on the intended design in as little as a few weeks. The prototype transforms written requirements into tangible experiences that foster alignment by removing ambiguity. This "show-don't-tell" approach of expressing requirements through design and interaction supports alignment and understanding of what we're building and how it works.

Iterative Design Refinement (MVP and Scale-Up)

Once Beta goes live and the focus shifts to MVP development, design becomes embedded within sprint teams, operating iteratively while continuously refining the experiential prototype as a "north star." Our design team works 2-3 sprints ahead of development, so planned features are discoverable, candidates for rapid low-fidelity experiments, and usable before they enter the development pipeline.

During this phase, our design strategist focuses on:

- **Conducting ongoing research** by collecting user feedback and observing usage metrics from released versions through both qualitative and quantitative methods, including focus groups, interviews, and analytic tools.
- **Synthesizing insights** with our established vision to continuously refine the product roadmap.
- **Building feedback loops** between site administration, usage analytics, customer support, design and marketing teams to understand real-world behaviors first-hand.

The User Interface (UI) Designer remains focused on incorporating those changes into screen-level designs, driving clarity for engineering teams and stakeholders.

Using this approach means we don't just deliver features, but build a platform that users will actively adopt, engage with, and continue using over time.

Design Deliverables

DELIVERABLE	DESCRIPTION
User Personas	Up to 5 research-based user personas representing the primary user groups identified in Section 4.0 of the requirements. Personas synthesize user research into archetypal profiles that capture goals, pain points, technical capabilities, and preferred workflows. Examples may include personas representing different levels of technical expertise (from AI researchers to

DELIVERABLE	DESCRIPTION
	civic organizations) or different use cases (dataset discovery vs. synthetic data generation vs. Commons Credits management).
User Journey Map	User journey map representing the end-to-end experience for personas across key stages and workflows, including but not limited to dataset discovery and access, Commons Credits earning and spending, synthetic data generation and fairness evaluation and bias detection. Map identifies touchpoints, pain points, and opportunities at each stage.
Experiment Designs and Execution	Atomizing the user journey into individual "What Must Be True" statements, the team creates hypotheses and null hypotheses that can be blind-tested with real candidate users to determine feasibility, desirability, economic viability, durability, and ethicality of various solution design elements and experience mechanics.
Interactive Prototype (3 versions)	Clickable prototype demonstrating user journeys and interactions across touchpoints. Includes the following system interactions: Commons Credits, Dataset Discovery & Filtering, Synthetic Data Generation, Fairness Evaluation Tools, Admin Dashboard Functionality. Validates complex interactions before development begins.
Information Architecture (Sitemap)	Comprehensive site structure, including navigational hierarchy, content taxonomy and categorization system that outlines how users will find and discover information and features.
Service Blueprint	Detailed mapping of front-stage user interactions, back-stage system processes, integration points between components (catalog, credits, synthetic tools), governance and approval workflows, and data flow between external datasets and hosted content. Identifies operational requirements and technical dependencies.
Design System & Component Library	Comprehensive design system including reusable UI components, interaction patterns, visual design standards, and technical specifications. Supports consistency across all platform touchpoints and accelerates development.
Production-Ready Designs	Detailed implementation documentation including layouts, component specifications, responsive behavior, interaction states, and styling guidelines. Provides developers with designs, asset libraries, and technical annotations needed to implement interfaces.

DELIVERABLE	DESCRIPTION
Enablement Strategy & Plan	Defines Enablement approach and activities across engagement, communications, learning, and metrics; plan defines frequency/cadence, ownership, and timing. Aligns to program timeline and key milestones.
Change Network Approach	A community of employees who work with the Enablement team to support and enable a successful journey for impacted users. Champions provide feedback to the program team and provide perspective on end user adoption tactics. Efforts include managing the network, creating content for recurring meetings, creating content for Champions to cascade, owning communication pre-reads and follow-ups, overseeing a shared collaboration space, etc.
Communications Collateral	Materials (e.g., forum posts, newsletters, office hours, walk around content) to build awareness, understanding, and buy-in of the DCC.
Self-Service Learning Materials	Support user learning and upskilling. Includes webinars, quick reference guides, videos, etc.
Consumer Insights report and Implications for Design	This is a living document that will be updated every 2 weeks. The framing for this report will be "what behavior did we observe? How does that behavior compare to what we hoped we would see in a Commons? How must users behave differently at scale to get the outcomes we want?" These questions will help us formulate our next set of hypotheses of "What rules do we need to implement to drive those behaviors?"

Technology Approach

We've designed a modern, cloud-first architecture centered around a unified user interface, API-based microservices to seamlessly integrate open source and best-of-breed commercial services, and flexible data management. Our guiding principles for the design center around:

- **Modularity |** Each component of the solution functions as a standalone service that communicates through standard interfaces.
- **Interoperability |** All components adopt open standards (DCAT, Schema.org, OIDC, OpenTelemetry) and expose APIs for integration. Metadata is structured to align with

FAIR principles (Findable, Accessible, Interoperable, Reusable), so that datasets are portable and usable across Commonwealth, academic, and industry platforms.

- **Security |** Security and compliance are built in from the start, with identity federation (OIDC/SAML), RBAC, MFA, encryption in transit and at rest, centralized audit logs, and compliant data residency.
- **Cost Realism |** The design favors proven open-source components and elastic cloud services, avoids lock-in, and right-sizes infrastructure to balance performance with budget considerations, supporting sustainability for the Commonwealth.

The core user experience is built on a custom portal that provides search, catalog, dashboards, data workbenches, and other features in a single, consistent format. This experience layer allows us to design end-to-end workflows in a unified environment despite multiple back-end tools. We envision a feature-based, iterative development approach to progressively make more functionality available to users over the course of the development process.

We believe this architectural approach aligns closely with the DCC's objectives of a hybrid cloud/MGHPCC architecture, open-source software (OSS) and standards-first components, and compliance with FAIR principles throughout the solution.

Architectural Overview

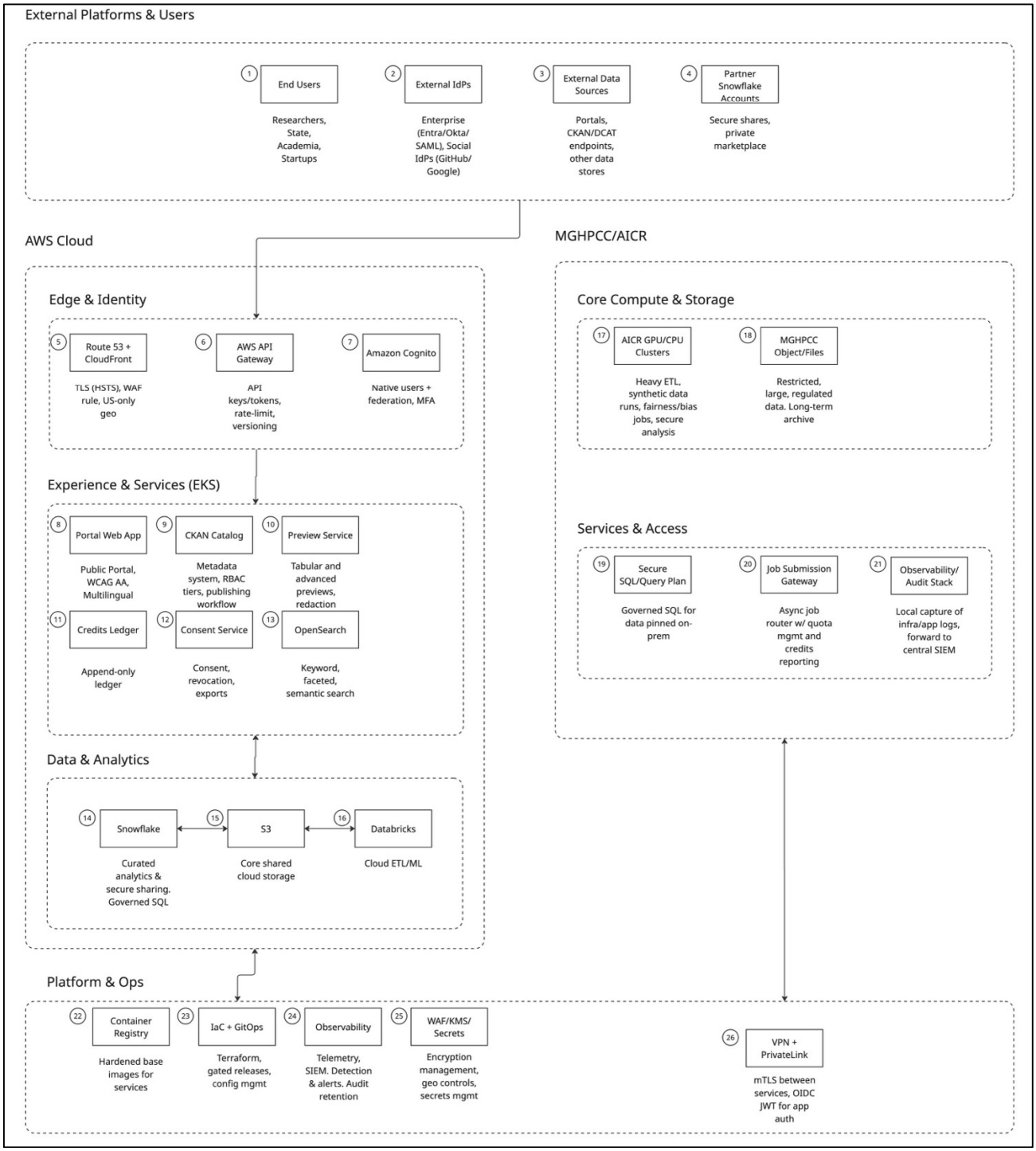

External Platforms & Users
1 End Users
Researchers, State, Academia, Startups
2 External IdPs
Enterprise (Entra/Okta/ SAML), Social IdPs (GitHub/ Google)
3 External Data Sources
Portals, CKAN/DCAT endpoints, other data stores
4 Partner Snowflake Accounts
Secure shares, private marketplace
AWS Cloud
Edge & Identity
5 Route 53 + CloudFront
TLS (HSTS), WAF rule, US-only geo
6 AWS API Gateway
API keys/tokens, rate-limit, versioning
7 Amazon Cognito
Native users + federation, MFA
Experience & Services (EKS)
8 Portal Web App
Public Portal, WCAG AA, Multilingual
9 CKAN Catalog
Metadata system, RBAC tiers, publishing workflow
10 Preview Service
Tabular and advanced previews, redaction
11 Credits Ledger
Append-only ledger
12 Consent Service
Consent, revocation, exports
13 OpenSearch
Keyword, faceted, semantic search
Data & Analytics
14 Snowflake
Curated analytics & secure sharing. Governed SQL
15 S3
Core shared cloud storage
16 Databricks
Cloud ETL/ML
Platform & Ops
22 Container Registry
Hardened base images for services
23 IaC + GitOps
Terraform, gated releases, config mgmt
24 Observability
Telemetry, SIEM. Detection & alerts. Audit retention
25 WAF/KMS/ Secrets
Encryption management, geo controls, secrets mgmt
MGHPCC/AICR
Core Compute & Storage
17 AICR GPU/CPU Clusters
Heavy ETL, synthetic data runs, fairness/bias jobs, secure analysis
18 MGHPCC Object/Files
Restricted, large, regulated data. Long-term archive
Services & Access
19 Secure SQL/Query Plan
Governed SQL for data pinned on-prem
20 Job Submission Gateway
Async job router w/ quota mgmt and credits reporting
21 Observability/ Audit Stack
Local capture of infra/app logs, forward to central SIEM
26 VPN + PrivateLink
mTLS between services, OIDC JWT for app auth

Figure 11: Technical Architecture

We envision a microservices-based architecture with discrete, modular services for the core catalog, data ingestion, search, credits, synthetic data generation, fairness, and other supporting services. This approach allows for incremental delivery, easier service scalability, and independent evolution of the individual services to progressively enhance the platform over time with minimal complexity.

The hybrid nature of the architecture supports on-demand, easily scalable, flexible cloud services combined with the ability to seamlessly tap into the compute and storage capacity of MGHPCC/AICR to support sensitive or large-scale datasets. Combining all data assets (tabular and object-based, cloud and on-prem, internal and federated) into a single catalog gives users a simpler, more efficient experience. Tight integration with MGHPCC enables those organizations with compute or data resources housed there to easily meet compliance requirements and leverage their existing assets and capabilities.

Flexible, Modular, and Secure Architecture.
Our solution features a modular, cloud-first architecture built on open standards and best-of-breed tools. Each component operates as a standalone service, enabling rapid deployment, scalability, and seamless integration with state and external platforms. Security and compliance are embedded from the start, with robust identity management, encryption, and auditability. This design approach focuses on building a resilient, cost-effective, and future-ready Data Commons — supporting evolving needs while maintaining interoperability and high performance.

Cloud resources will be deployed in east-coast regions and optimized for elasticity and cost control. Public access will be governed by security best practices to reduce the attack surface and zero trust, API-based interactions between services help contain the blast radius of incidents that may occur.

Because the architecture is intentionally componentized, each service and functional area can use best-of-breed tools to deliver first-class functionality with smart cost controls and maximum flexibility. We plan to use proven OSS and commercial components like Kubernetes, Terraform, AWS Secrets Manager, CKAN, OpenLineage, Elasticsearch, AWS API Gateway, Snowflake, and Databricks to allow us to strike the right balance between capabilities, implementation effort, ongoing operations and support needs, and ongoing cost. Our API-first approach and common metadata standards support changing individual components without wholesale refactoring to respond to evolving capabilities and costs in the future.

By contrast, a single, all-in-one platform would include unnecessary features, limit flexibility, and still require additional integrations to meet real-world needs. Our modular approach avoids that trap. It gives MassTech the freedom to evolve and swap components as technologies and requirements change—an area where Slalom's deep experience in system integration enables seamless interoperability and sustained performance.

The sections that follow describe several key components that make this architecture both flexible and practical.

Core Catalog

We propose CKAN as the backbone for dataset cataloging and metadata management. CKAN provides full API coverage for its features, enabling tight integration with the custom portal front end as well as other services in the solution. CKAN's robust extension capability allows us to implement core features like DCAT/Schema.org support and secure previews with minimal development overhead. A rich ecosystem of CKAN plugins allows for configuration instead of custom development for core requirements. The flexibility of the extension system allows us to incorporate features like governance and approval workflows or credit consumption without standing up separate services. It also allows us to more tightly integrate features like synthetic generation or bias detection into the core catalog functionality.

Ultimately, CKAN will act as the authoritative system of record for dataset metadata, provenance, lineage, and access controls.

DevSecOps Foundation

Our approach heavily leverages modern automated development and security processes. Infrastructure as Code tools, such as Terraform, and GitOps tools, such as Argo, will be used to automate the deployment and management of components of the application. This allows version control and review/acceptance gates for core configuration, simplifying the configuration management process and removing process bottlenecks that can arise in more manual approaches.

Smart CI/CD pipelines and processes will support comprehensive automated testing, automatic security scans, and hardened base images for individual services. The initial investment of development effort early in the work pays back dividends each time teams check in code.

Cloud Data Platforms

We propose using both Snowflake and Databricks as core parts of the solution. These tools overlap superficially, but together they make the system more capable and flexible than either one alone. Each contributes a specific set of capabilities and serves discrete functions in the overall architecture.

Both platforms act as pluggable accelerators; the catalog remains the system of record for metadata and access policies with specific data serving, analysis, manipulation, and compute needs being handled by the two platforms. Having both tools available opens up the opportunity to select per-workload based on cost, governance needs, and required capabilities and avoids single-vendor lock-in for core components of the application.

Better together—Snowflake & Databricks.
Snowflake is optimized for curated, governed publishing and secure external sharing of datasets once ingested and approved. Databricks is optimized for heavy data engineering, ETL of cloud datasets, and compute-intensive workloads outside MGHPCC. Together they provide complementary capabilities, made more useful by the unified catalog that is central to the solution.

Snowflake. We envision Snowflake acting as the curated analytics and sharing layer for published datasets that do not have security or residency requirements that cause them to be hosted on-prem. Snowflake's secure data sharing capabilities, combined with the significant Snowflake footprint across Commonwealth agencies, will allow more seamless access and serving of high-value datasets without incurring additional storage and processing costs to actually ingest them. Secure data sharing will also allow capabilities like sharing data from the platform with users who already have Snowflake as part of their data infrastructure without incurring additional data egress or processing costs.

Snowflake can also be used as an efficient, simple-to-manage platform for data and analytics internal to the solution. As an example, the secure managed environment can support building and querying data marts focused on telemetry, UI clickstream, and data consumption details that will enable valuable insight into how the solution is performing and how users are interacting with it.

Longer term, Snowflake can unlock opportunities for providing SQL analysis capabilities to users, including data stored in Snowflake as well as external tables configured over lake storage. The Commonwealth can also take advantage of the private marketplace feature to share and consume datasets between departments or secretariats, allowing an easy way to incorporate data from the DCC into existing workflows or develop novel solutions.

Databricks. We envision Databricks as the cloud ETL engine for datasets ingested and processed outside MGHPCC. Slalom has deep expertise and many successful experiences implementing efficient, templatized, metadata-driven data pipelines and has collaborated with Databricks to develop accelerators that have been used across various industries and use cases. Modern catalog standards like Iceberg further simplify using Databricks compute with datasets managed in Snowflake. This allows us to avoid having to make tradeoffs between efficient, cost-effective cloud compute and simple, easy to use management and querying. Databricks' inbuilt metadata and lineage functionality through Unity Catalog allow easy integration with the CKAN catalog without requiring complex orchestration or additional tooling.

Additionally, Databricks presents an opportunity to provide scalable, on-demand compute resources to users. Features like MLflow for experiment tracking and model registry would support responsible ML development patterns for organizations that may not have in-house capabilities and provide another avenue to use data hosted within the solution without needing to export and download datasets.

Security and Compliance by Design

We anticipate a varied blend of user authentication models based on the information in the RFP and our understanding of the user base. We envision a tiered approach to identity and access management, with the foundation being Amazon Cognito as the application's identity broker. Cognito User Pools will support native identity management for users not authenticating with an external IdP. Cognito federation will allow SAML or OIDC federation with Entra, Okta, or other enterprise SSO providers for organizations that choose to federate. Cognito social federation could, if desired, allow sign-in using other providers such as Google or GitHub accounts.

By having Cognito act as the central identity broker, we can achieve a single brokered token format (i.e., Cognito-minted OIDC JWTs) regardless of how a user logged in, so the API gateway/ALB authorizers and backend services implement one unified verification path. MFA enforcement at login and at the app layer support security controls for specific datasets or sensitivity tiers.

Data protection is treated as a first-class priority of the solution, with comprehensive encryption in transit and at rest throughout the solution further secured by a zero-trust architecture. In a zero-trust environment, every user and every service must authenticate and be authorized on each call. Aside from JWTs vended by Cognito, services will use mutual TLS--both parties verify the other's identity using centrally managed certificates before they can interact. JWTs and the RBAC scheme combine to allow enforcement of access restrictions at the dataset and API endpoint level.

Centralized logging of all catalog, API, and platform interactions with long-term retention will provide auditability and the ability to analyze usage patterns in aggregate. Logs, metrics, and traces will be integrated into a centralized SIEM for monitoring and incident response. Exportable audit reports will support compliance reviews and, where appropriate, research reproducibility.

The DevSecOps processes described above will include automated accessibility compliance checks as well as automatic security testing to provide an additional layer of defense against vulnerabilities.

Observability & Resilience

The solution will be designed with logs, metrics, and traces for all services using OpenTelemetry-compatible agents to maximize interoperability. Standard correlation IDs

across the services and front end will allow traceability of end-to-end user journeys, and feeds will be collected in a central SIEM/observability stack for auditability.

Privacy-first product analytics will be implemented to provide valuable insights into user behavior and accessibility of various features. Clickstream analytics will be carefully constructed to avoid capturing PII and will honor users' Do-Not-Track preferences. Pseudonymous sessions with configurable retention windows will help effectively separate identifiable information from valuable behavioral data that will help enhance the usability and utility of the solution.

Operational, governance, product, and security analytics with accompanying dashboards will provide a comprehensive picture of key metrics and health indicators like latency, error rates, ingestion throughput, dataset views, authentication failures, and permission denials. If appropriate, given the overall threat model, the underlying telemetry data can also be incorporated into advanced security tools (commercial or open source). This telemetry approach will also enable robust, configurable alerting and incident response that can trigger varying levels of escalation for intervention.

Observability also extends to the cost and usage areas, including per-service cost attribution via tags and labels. Usage tagging can also extend to compute and storage within Snowflake and Databricks, enabling capabilities like per-tenant accounting that could be incorporated into the Commons Credits ecosystem. At a solution level, economic KPIs like cost per 1k searches, cost per preview, or cost per 1M catalog API calls could be used to prioritize investments tied to cost reduction.

A multi-availability zone deployment for the cloud workloads, combined with health-checked auto-scaling and rolling updates will support resilience throughout the stack. Backups and point-in-time recovery for databases and object storage indices will reduce the impact of service interruptions. The Infrastructure as Code and GitOps strategies described earlier further enhance resiliency by allowing for both automated, self-healing and rapid intervention by operational teams.

Build Approach

We strongly recommend adopting an iterative product development approach that delivers working software early, gathers real user feedback, and uses that feedback to guide each next step. Observed behavior and input from early users provide critical insight, helping us focus investment where it will deliver the most value. The architecture described above is designed to support this approach.

We envision a phased approach to building the technology components of the overall solution. The **Foundation and Beta phase** establishes essential infrastructure—EKS clusters, networking, security baselines, API gateway, catalog service configuration—and an initial version of the user-facing experience. These are "no-regret" investments that remain valuable regardless of how user needs evolve.

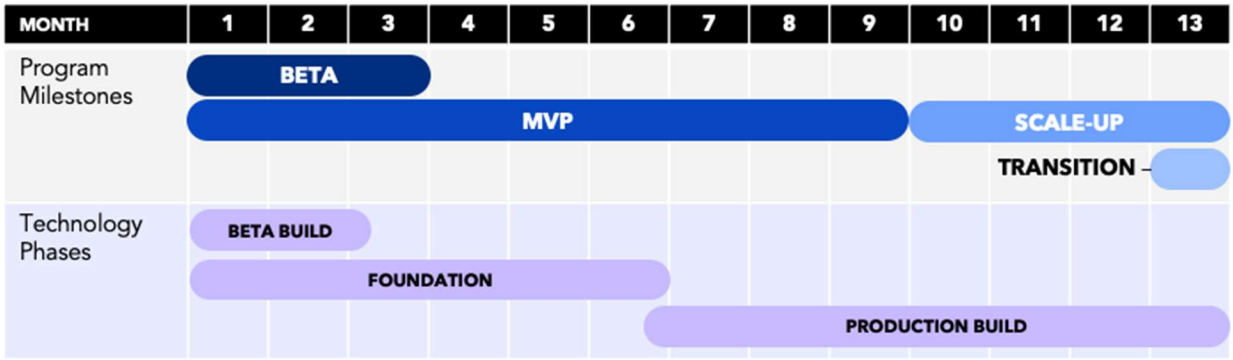


Figure 12: Project Timeline With Technology Phases

We plan to develop a more robust initial public release (MVP) in parallel with development and validation of the Beta feature set. This allows us to shorten the overall duration of the program while incorporating insights from Beta to guide MVP. Continuous feedback and iteration keep the product flexible, inform release planning, and reduce the risks inherent in building a new platform with novel capabilities.

Continuity & Handoff

This architectural approach is focused on building a sustainable product that will be extensible, manageable, and cost-effective. Throughout the development process, comprehensive documentation at the code and solution levels will be created and maintained, and focused training will be provided for the teams that will ultimately own and maintain the product after this initial development is completed.

Sustainability extends to a focus on creating a high-quality, modular codebase with low technical debt. The architecture is designed for extensibility, both through the selection of extensible tools like CKAN as well as the services-based design that allows additions and enhancements of components without large rewrites.

This approach culminates in the development of a clear ownership transition plan at the close of the engagement. Our teams will focus on enabling MassTech to both manage and continue the enhancement of the solution. Simply put, our goal is to make ourselves obsolete.

Adoption Approach

Technology only creates value when people use it—and keep using it. Our Adoption team helps the Data Commons attract, engage, and sustain a vibrant user community through communications, learning, and continuous feedback.

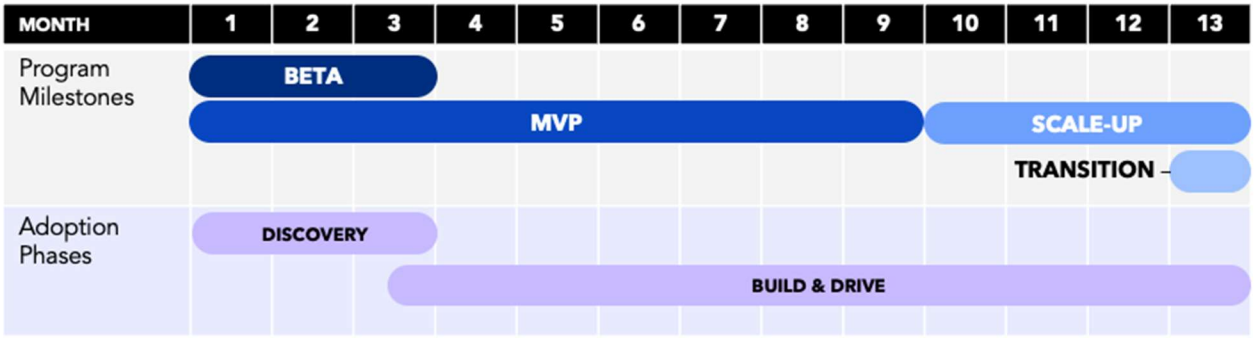


Figure 13: Project Timeline With Adoption Phases

User Discovery & Engagement Design

This phase lays the foundation for sustained adoption by focusing on understanding user needs and defining scalable engagement strategies. We will:

- Conduct **interviews** with key MassTech decision-makers and other stakeholders to understand use cases, challenges, motivators, current ways of operating, learning needs, and preferred communications and learning channels. This informs the design of end user outreach and communications, user onboarding and learning pathways.
- Define **Adoption KPIs** that are aligned with MassTech's definition of success. This could include number of users, daily active users, percentage of users referred by other users, dataset count, usage per dataset, number of credits transactions, etc.
- Develop a comprehensive, persona-based **Adoption Strategy and Plan** informed by early practical experiment results and stakeholder interviews that integrates communications, engagement, learning, and metrics. This plan will guide how users are introduced to the DCC and supported over time.
- Begin designing **outreach campaigns** and the approach to standing up a **champion network**, including public, private, and academic sector representation.
- Design and launch **foundational messaging** to build awareness of the DCC's purpose, capabilities, upcoming pilot opportunities, and avenues to learn more.

Engaging Users for Enduring Adoption.
Our adoption approach helps the Data Commons attract, engage, and sustain a vibrant user community. We start by understanding user needs and designing scalable engagement strategies, including targeted communications, onboarding, and learning resources. Continuous feedback and metrics guide ongoing improvements, while champion networks and outreach campaigns foster peer-to-peer learning and long-term platform success. This strategy creates awareness, drives meaningful use, and supports users throughout their journey.

- Establish **forums** to foster peer-to-peer learning and engagement. This includes defining a strategy to **recognize and reward** valuable contributors through the Commons Credit system.

Figure 14: Illustrative Discovery & Design Phase Activities

This phase grounds our adoption strategy in real user needs and sets the stage for long-term success.

During the build phase, our focus shifts from planning to execution—continuously engaging users, sharing progress, and refining content and tools based on feedback. We will:

- Leverage identified communications channels, including forums, to **distribute key messages** and updates to the user population. We plan to design and deliver communications to end users before each phased release (Beta, MVP, At-Scale) to create awareness and excitement, as well as ongoing communications to keep up engagement.
- Deliver structured **onboarding** through webinars, self-service resources, and optional office hours for support.
- Build and iterate on **learning materials** tailored to end user personas and varying degrees of technical proficiency; includes webinar presentations, reference guides, and videos. We will then **facilitate learning** through webinars and driving access to self-service resources (e.g., reference guides, videos); topics may include data access, synthetic data generation, fairness dashboards, and data maintenance.
- Implement mechanisms for **continuous feedback** (e.g., support tickets, champion network, community input) and track onboarding, **adoption, and usage metrics** (e.g., dataset downloads, Commons Credit transactions).

Figure 15: Illustrative Build Phase Activities

Adoption is about the "push"—creating awareness and motivation—while great user experience creates the "pull," making people want to return and contribute. These two teams work harmoniously to apply the right tactic for each situation. Driving adoption is a dynamic process of listening to user needs, learning, and adapting so that users are supported throughout their journey with the DCC.

Organizational Enablement Approach

Launching the Data Commons is not just a technology build—it is closer to starting a new business. A platform like this needs an organization capable of running it. Our Organizational Enablement work helps MassTech design the structures, roles, and governance needed to operate and evolve the Data Commons for years to come. Beyond code and infrastructure, the platform will need clear governance processes, day-to-day administration, support channels, and mechanisms for marketing and continuous improvement. MassTech has not yet specified how these pieces should work, so our role is to help surface any blind spots and co-design a sustainable model.

To guide this work, we use the Mission Model Canvas (MMC)—a proven tool for mapping out all the moving parts of a mission-driven initiative. The MMC forces visibility into elements often missed when

Strategic, Sustainable, and User-Aligned Enablement

Our organizational enablement approach helps MassTech design the structures, roles, and governance needed to operate and evolve the Data Commons for years to come. Using proven tools like the Mission Model Canvas, we co-design a future-state operating model that enables clear accountability, robust support channels, and continuous improvement. This work aligns governance and operations directly to real user needs, reducing risk and building a platform that is technically robust, ethically grounded, and institutionally sustainable.

thinking purely as a product build—such as customer support, marketing, and funding models. By doing this early, we avoid costly surprises and tie governance and operational design directly to real user needs.

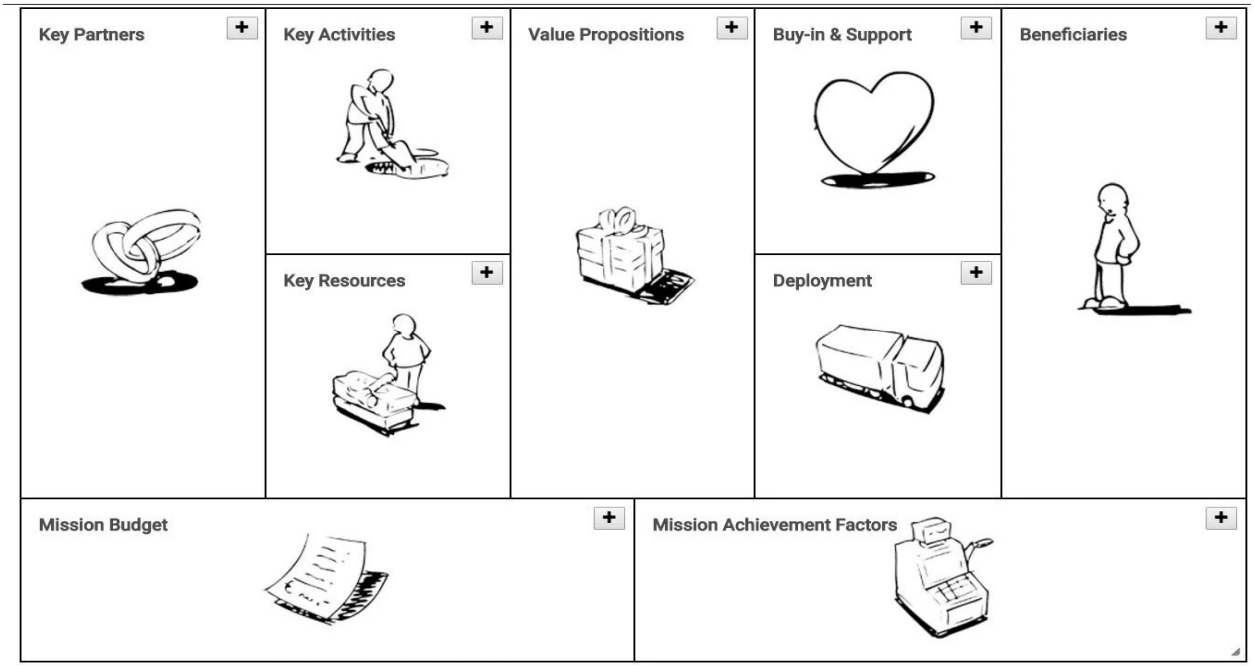


Figure 16: Example of a Mission Model Canvas

Strategize & Design

This work is focused on defining the foundational elements of the operating model and preparing for implementation. We will:

1. Conduct a half- or full-day workshop with stakeholders using the Mission Model Canvas to make sure all stakeholders are aligned about the desired outcomes, measurements, funding mechanism, and other components needed to run the future-state org. This will serve as our expected end-state to work towards.
2. Map the future-state organizational structure that will own the DCC, identifying the teams, roles & responsibilities, and interdependencies required to operate key functions (data governance, user support, platform maintenance, outreach).
3. Design a **tech maintenance governance structure** for after the Slalom development team rolls off, enabling cross-functional collaboration and clear accountability. Develop a comprehensive plan for go-forward project phasing, resource management, and budgeting.
4. Define **key operational processes:** intake workflows (user submission & approval, role assignment), data quality and sensitivity checks, quality assurance, site performance, catalog management, and Commons Credits reward mechanism. This would also consider any workflows associated with the Admin UI.

5. Partner with institutions to **co-design forums** that address policies, data usage, quality standards, privacy protocols, and collaborative data sharing.
6. Consider **non-product operational components** such as call center support, system administration, marketing, finance / accounting and business development.
7. Plan for **tiered support channels** to address user needs, including escalation paths. Identify a portal-based system (e.g., ServiceNow[1]) for troubleshooting and inquiries if justified.

This operating model enables DCC to be not only technically robust, but also ethically grounded, user-responsive, and institutionally sustainable.

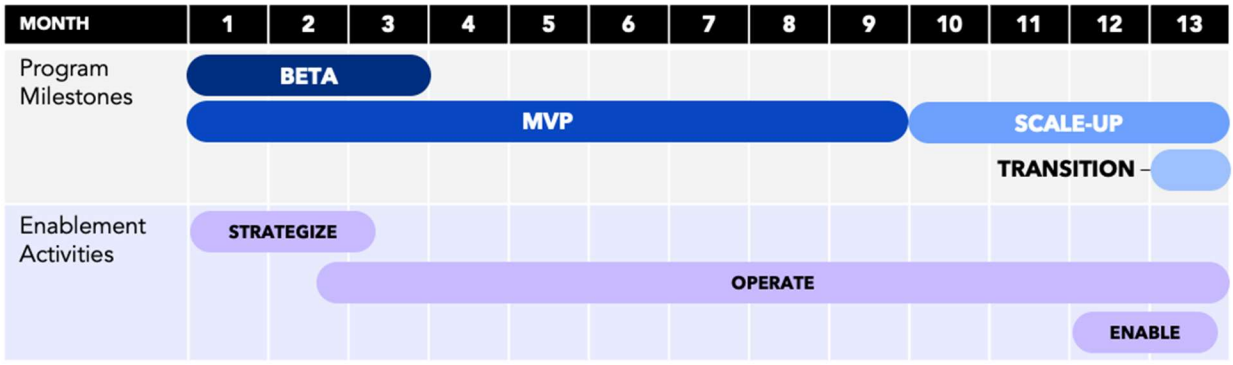


Figure 17: Project Timelines With Enablement Activities

With the overall approach defined, the following sections provide a closer look at four technical pillars that make the Data Commons unique and cross-thread through the workstreams defined above. Each represents a core capability required for launch and scale—integral to both the technology architecture and the user experience.

Key Topic: Data Catalog

Introduction

The Data Catalog is the backbone of the Data Commons Collaborative. It is the central entry point for discovering, understanding, and responsibly accessing datasets across cloud services and the MGHPCC/AICR environment. We propose **CKAN** as the core catalog platform, leveraging its open-source extensibility, mature plugin ecosystem, and full API coverage to meet the requirements of FAIR data management, security, and interoperability.

[1] Note: Our intended scope of work does not currently include implementing a customer service platform. We can envision future scenarios where a robust service function would be very useful, and other scenarios where it is not needed.

The catalog lowers the barrier to dataset access by offering a unified, user-friendly directory of datasets that are curated, discoverable, and traceable. It enables:

- **Discovery** – Searchable metadata aligned with DCAT/Schema.org and FAIR principles, enriched with provenance and sensitivity indicators.
- **Responsible Use** – Clear licensing and tiering (Public, Restricted, Internal) provide appropriate governance for dataset access.
- **Transparency** – Lineage and versioning records establish trust and reproducibility.
- **Collaboration** – A single system of record for metadata connects public, academic, and industry contributors, linking datasets to synthetic derivatives and Commons Credits tracking.

Core Capabilities

The catalog's contribution to the overall solution is focused around four core capabilities.

- **Dataset Directory |** A consolidated web-based entry point presenting dataset "cards" with persistent identifiers, metadata, and links to APIs or endpoints. Datasets are clearly marked by access tier, guiding users to the right dataset for their use case.
- **Metadata Standards |** Adoption of a DCAT/Schema.org profile supports interoperability and enables discovery. Required fields will include title, description, owner, license, modality, sensitivity, provenance, version, previews, endpoints, and keywords. The schema is extensible to add quality scores, bias analysis, and FAIRness metrics over time.
- **Search & Discovery |** Search is powered by both keyword indexing and semantic embeddings for relevance beyond keywords. Users can refine results with faceted filters (e.g., domain, license, or sensitivity).
- **Dataset Previews |** Secure previews (tabular, text, and image) help users assess dataset relevance without downloading full files. Guardrails enforce privacy so that sensitive data is not exposed.

Data Lineage & Provenance

The catalog records dataset-level lineage at launch, capturing ingestion source, transformations, and derivations, with a roadmap to finer-grained column-level lineage. Provenance metadata includes contributor, submission date, and rationale for changes. Synthetic datasets are explicitly linked to their original sources with validation metrics included.

Data Ingestion & Curation

We will implement batch-first ingestion pipelines optimized for bulk uploads and initial curation. Automated metadata validation and profiling will run at ingestion time,

supplemented by curator dashboards for review, approval, and enrichment. Contributors will submit datasets via structured forms aligned to the DCAT profile with differentiated upload paths: MGHPCC for high sensitivity/large-scale data, and cloud for lighter public datasets. Contributors will also be able to submit externally hosted datasets, allowing additional breadth and variety of data available to users without incurring additional storage and serving cost for each.

Governance & Access Control

The catalog enforces Role-Based Access Control (RBAC) and dataset tiering with SSO and MFA provided by upstream services. Fine-grained access permissions are tied to catalog objects like datasets, metadata, and previews with roles defined for contributors, reviewers, admins, and external users. We anticipate an initial three-tier sensitivity model, but this can be refined to support Commonwealth standards or DCC-specific requirements. Centralized logging of all catalog interactions will be retained for at least 3 years to support security and auditability requirements. Integration with Security Information and Event Management (SIEM) supports robust monitoring, alerting, and incident response capabilities. Audit reports can be exported to support compliance and reproducibility.

Integration & Interoperability

A core requirement of the Data Commons is that the catalog must not exist in isolation. It must seamlessly connect datasets across MGHPCC, cloud platforms, and external registries while enabling security, compliance, and interoperability. Our design embeds integration at three levels: infrastructure, application, and standards.

MGHPCC/AICR Connectivity

The catalog will integrate directly with MGHPCC and the state's Artificial Intelligence Compute Resource (AICR). High-volume, high-sensitivity datasets can be ingested or staged on MGHPCC infrastructure, while metadata and access pathways are registered in the catalog. This hybrid architecture keeps metadata and APIs cloud-accessible, while restricted workloads run within MGHPCC's secure, controlled environment. This ultimately supports the Data Commons requirement for data residency, sovereignty, and tiered access, while leveraging MGHPCC's AI cluster for compute-intensive operations like synthetic data generation or fairness audits.

API-First Design

All catalog features (search, submission, metadata retrieval, lineage queries) are exposed via REST and, if desired, GraphQL APIs. API parity allows users to perform any action available in the portal UI programmatically, enabling integration with external research tools, Juypter notebooks, or teaching platforms. APIs will support rate limiting, quotas, and API key/token management to align with the DCC's multi-tenant security model. Over time, advanced APIs

may allow federated search across external catalogs, automated harvesting of CKAN/Dataverse sources, and audit log exports for compliance review.

Standards-Based Exchange

Metadata export formats include DCAT, JSON-LD, and CSV, supporting compatibility with national and global open data catalogs. Support for persistent identifiers (DOIs/DataCite) promotes dataset “cite-ability" and reproducibility. Linked data vocabularies (DCAT/Schema.org/Dublin Core) make the catalog interoperable with FAIR data infrastructures across academia, government, and industry. Provenance and lineage records will be aligned with open standards like OpenLineage to support reproducibility and governance reporting, as well as clear user understanding of the provenance of data in the solution.

Interoperability with State Platforms

The catalog will interoperate with state data and analytics platforms, including Snowflake (already in use by many state agencies) and other relevant systems. Integration pathways will be designed to support future partnerships with the Mass Open Cloud, healthcare research infrastructures, and regional data commons initiatives. By exposing standards-based APIs, the catalog becomes a hub where agency data, MGHPCC compute, and commercial platforms can be securely bridged without vendor lock-in.

Observability & Metrics

The solution will capture the minimum viable user telemetry to prove adoption, enforce governance, and inform continued product development—all without collecting PII. This aligns to DCC requirements for usage metrics in the catalog, RBAC/auditability, SIEM forwarding, and retention. Examples of signals that will be collected include:

- **Events |** Dataset views, searches, preview opens/sizes, downloads, request-access submissions, or curator actions. Users’ Do Not Track preferences will be honored with opt-out testing.
- **Audit Logs |** Authentication (“Who are you?”) and authorization ("What are you allowed to do?”), dataset tier checks, approvals/denials, data exports, and synthetic run invocations.
- **Metrics |** Platform health metrics like API latency/error rates, search response times, ingestion throughput/success rate, queue depth, index size, or crawler/harvester status.
- **Traces |** Cross-service request paths (for example, portal -> APIs -> search -> storage) captured and stitched together to support root-cause analysis and service level objective reporting.
- **Cost & Usage |** Compute, storage, data egress, and API usage tagged by tenant and ready to be consumed by the Commons Credits ledger and product dashboards.

These signals will allow creation of management and reporting dashboards covering use cases like adoption and search quality, catalog and curation health, ingestion and data quality, and governance and audit events.

Accessibility & User Experience

The catalog is designed for inclusivity and ease of use from day one. We will build to WCAG 2.1 AA standards, with color contrast, keyboard navigation, and screen-reader compatibility being part of the acceptance criteria for every UI feature. Accessibility checks will be automated in CI/CD pipelines and validated with manual testing to reduce the risk of regressions.

Performance standards are equally important. The catalog will be optimized for lightweight, responsive operation across desktops, tablets, and mobile devices, with a target of P95 < 2 seconds load time for dataset searches and detail pages on reference datasets. Server-side rendering and caching will support usability even on lower-bandwidth connections.

Multilingual scaffolding will be built in for the initial public release, with English prioritized at launch but with an architecture ready for localization and translation expansion as the user base grows. Metadata fields will be designed for multilingual entry so contributors can supply descriptions in more than one language where appropriate.

To improve adoption and reduce training burden, the catalog will provide intuitive navigation and inline support. Clear navigation and breadcrumbing on search and detail pages help users know where they are. Tooltips and contextual guidance embedded in forms, such as metadata submission or request-access workflows, put valuable information and instructions where they're needed. Onboarding tutorials and guided walkthroughs for first-time users, designed with both technical and non-technical users in mind, will support quick time-to-comfort for new users. A consistent look, feel, and language across the portal and APIs enables learning one part of the system to apply everywhere.

Key Topic: Commons Credits System

This system operationalizes one of the Data Commons' core design goals: creating a self-sustaining ecosystem that grows stronger the more it's used. The Credits System is critical to preventing over-consumption and under-contribution.

Executed well, it amplifies the network effect of the Data Commons. Executed poorly, it *could* create barriers to adoption by adding unintended complexity. Slalom recommends a "layered" approach to designing and building the Credits System through which we capture the benefits and avoid the risks.

1. Our Approach: Start Simple, Layer Over Time

We propose a "Layering Strategy" for the Commons Credits System. At launch, the rules will be intentionally simple — a novice user can learn them in a minute or two. This lowers barriers to adoption, avoids over-building, and makes the system usable from day one.

The Layering Strategy follows the same principle as our overall phased approach: address uncertainty early and cheaply, then scale what works. By observing real user behavior before adding complexity, we avoid over-building or deterring adoption.

The initial ruleset rewards joining, logging in, contributing datasets, and inviting others—this is aligned with what matters most when launching a new marketplace product: attracting users and driving beneficial usage. It allows users to spend credits on premium data or compute, and it includes an option to request credits for projects. Balances and histories are visible in a central ledger. Credits are non-refundable, and any account suspected of misuse may be frozen for review.

After around 2-3 months of real-world use, we will introduce additional layers, such as:

- SPEND:
    - Differentiated pricing by dataset: bigger data sets cost more
    - More compute-intensive analysis and generation costs more
    - Efficiency / sustainability incentives (e.g., cheaper off-peak compute, green computer)
- EARN:
    - Quality-based rewards: bigger and better dataset contributions earn more
    - Social rewards: get credits for reviewing datasets, authoring instructional content, etc.
    - "Need-based" rewards: get credits for contributing data that is aligned with current priorities (e.g., "Robotics Dataset Month" promotions)
- GRANT:
    - Ability to apply for a credit grant for a specific project
    - Explore real-currency purchase of credits down the line (initially out of scope)

What is critical to success is that "layering" is informed by *real-world behavior* over assumptions.

2. Governance and Fairness

The RFP requires governance and transparency; our design delivers this without unnecessary complexity. We will:

- Implement a secure, auditable central ledger. Our recommendation is to avoid using blockchain for the foreseeable future. With proper rules governing data uploads and user permissions, this will be a platform where the actors can be trusted and blockchain will not be needed.
- Provide every user with a balance view, transaction history, and clear ways to spend and earn credits.
- Enable administrators to grant/remove credits and freeze accounts as needed.
- Publish clear credit policies to support consistent treatment across stakeholders.

This balance of simplicity and oversight incentivizes the positive behaviors that maximize the value of the Data Commons while minimizing risk of abuse and administrative burden.

3. Design and Validation Plan

Our delivery plan follows a sequence that builds confidence before investing in code.

As long as possible before implementation, we plan to test the incentive rules through a **tabletop simulation** with diverse stakeholders. Using cards and paper credits, participants will "play" the system: signing up, earning, spending, and requesting credits. This low-cost exercise helps confirm that the rules are understandable, identifies loopholes, and captures feedback before the heavy lifting of UI design and tech build. The simulation also engenders shared ownership through co-creation, increasing trust and adoption when the platform launches.

1. **During the Beta Phase we will:**
    a. Run a <½ day workshop to co-develop testable hypotheses about usage and credit-related behaviors. E.g., "we believe no one will run out of credits in 3 months," "we believe >25% of users will contribute at least 1 dataset that at least 1 other user will use in the first 3 months." This will include a plan for how we determine if our hypotheses are validated by real usage.
    b. Run a workshop with you to develop an initial ruleset that turns these hypotheses into rules (see Exhibit for an Example of how specific hypotheses can be converted into rules)
    c. Run tabletop tests internally to get an early indication of how well our proposed rules / incentives drive simulated behavior.
    d. We will incorporate feedback from the simulation into wireframe UI designs to be incorporated into an early release of the product (Beta or MVP). These screens will sit on top of a simple but secure ledger to track all transactions.
    e. Observe private beta users in real-time, periodically reaching out to them to get qualitative feedback and share prototypes of future iterations of Credit rules
2. **Design & Build MVP Enhancements:**
    a. Leverage the insights from the Beta and behavioral hypotheses to settle on a final set of Rules and UI designs for MVP (public release).
    b. Determine if there are additional features required for either the credits UI or the back-end ledger.
3. **Layer 2 Enhancements (Post-MVP, Scale-Up, At-Scale):** After MVP Launch, we expect to release new enhancements every 2 weeks in line with our product roadmap.

This staged approach helps the Credits System launch quickly, prove its value, and then mature alongside the broader Data Commons.

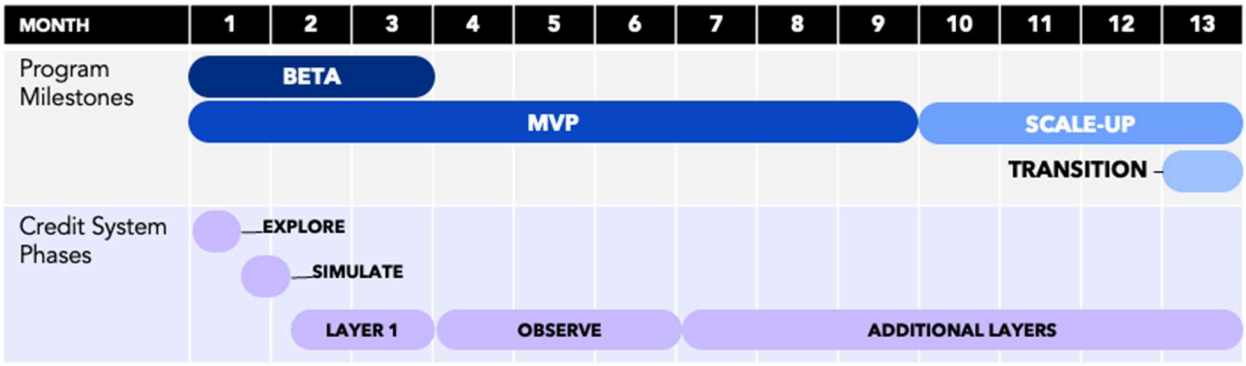


Figure 18: Project Timelines With Credit System Phases

5. Alternate Approach

While the Commons Credits System is our recommended solution, we recognize it is one of several proven ways to prevent the "tragedy of the commons." Other mechanisms—such as quota systems, social reputation, and tiered access governed by community norms—have succeeded in similar shared-resource environments.

To confirm that the chosen model fits Massachusetts' unique research and innovation culture, we propose a **two-week validation sprint** at project start. During this sprint, we will test the Credits approach against alternatives using brief simulations and stakeholder feedback. The goal is not to reopen design debates, but to **confirm that the Credits model is the most elegant, fairest, and most motivating mechanism** for this context.

If testing confirms that the Credits System best balances transparency, sustainability, and ease of use—as we expect—it will advance to detailed design. If a hybrid or variant model emerges as clearly superior, we will document the rationale and adjust accordingly in partnership with MassTech.

This structured validation step adds minimal cost but verifies that the final system is evidence-based, equitable, and fully owned by stakeholders from day one.

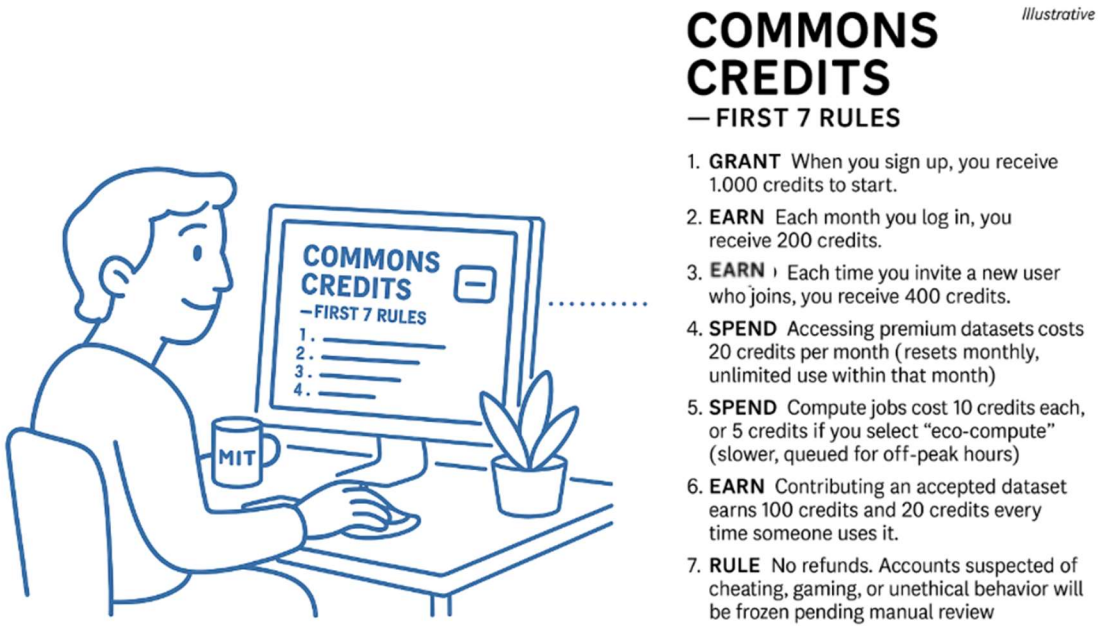
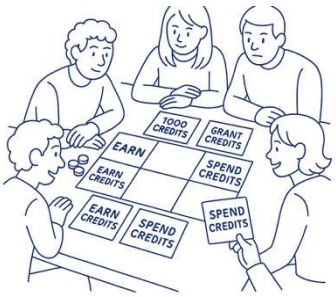

COMMONS CREDITS
—FIRST 7 RULES
1.
2.
3.
4.
MIT
Illustrative
COMMONS CREDITS
— FIRST 7 RULES
1. GRANT When you sign up, you receive 1.000 credits to start.
2. EARN Each month you log in, you receive 200 credits.
3. EARN Each time you invite a new user who joins, you receive 400 credits.
4. SPEND Accessing premium datasets costs 20 credits per month (resets monthly, unlimited use within that month)
5. SPEND Compute jobs cost 10 credits each, or 5 credits if you select "eco-compute" (slower, queued for off-peak hours)
6. EARN Contributing an accepted dataset earns 100 credits and 20 credits every time someone uses it.
7. RULE No refunds. Accounts suspected of cheating, gaming, or unethical behavior will be frozen pending manual review

Figure 19: Illustrative Ruleset for Testing in MVP Based on Behavioral Hypotheses

This ruleset is based on the following goals and hypotheses. These are examples that illustrate the method, not necessarily prescriptive recommendations. We plan to work with you to understand *your* goals and hypotheses.

- **Principle:** At first launch, the most important behavior to reward is logging in. The second most important behavior is bringing other users to the platform.
- **Principle:** Providing a healthy up-front credit balance allows users to engage with the system without pressure to contribute right away. But the balance shouldn't be so big that they use it "worry-free" forever. We want users to contribute in their first 3 months.
- **Principle:** We want to seed responsible / sustainable compute usage, and frame it as rewarding a positive choice, as opposed to penalizing a negative one.
- **Hypothesis:** Users will, on average, consume ~10x as much as they contribute.
- **Hypothesis:** Users will want access to 1-3 premium data sets per month, and it will likely be the same data sets over and over.
- **Hypothesis:** We believe 80% of user credit spend will be on compute. Either through creating synthetic data, analyzing existing data, or training models.
- **Hypothesis:** We expect users to run ~30 compute jobs / month.

Tabletop gaming is a low-cost way to simulate end-user behavior and explore multi-variate outcomes before investing resources and time into product development. Simultaneously, it allows us to bring stakeholders along into the decision-making process and garner excitement, buy-in, and momentum.



Key Topic: Fairness Tooling

The Fairness Tooling capability is a core component of the DCC's Responsible AI framework, so that all published and synthetic datasets meet equity, transparency, and trust standards.

Many organizations and AI developers increasingly rely on synthetic data to supplement or replace real datasets. Under such augmentation needs, the question of fairness becomes critical. Synthetic data must not only reflect statistical fidelity to source distributions, but also avoid amplifying biases, under-representing populations, or misrepresenting sensitive groups.

The Fairness Tooling service provides the framework to evaluate, monitor and improve the equity of datasets before they are shared, downloaded or used for downstream purposes. This section outlines the principles, capabilities and implementation approaches we can bring to construct a fairness checker which can operate independently or in concert with synthetic data generators.

Why Fairness Checking Matters: Fairness issues arise because real-world data often reflects historic or systematic bias. If such patterns are reproduced without intervention, synthetic datasets may continue to perpetuate inequity. Examples include:

- **Financial services:** Loan approval data that disadvantages women or minority borrowers.
- **Healthcare:** Clinical records that underrepresent certain ethnic groups or age ranges.
- **Autonomous Driving:** Point cloud datasets dominated by clear weather, daytime conditions with little adverse weather scenarios. Note: while not "unfair" per se, this is an important example of bias that would impede research and innovation

Transparent, Auditable, and Domain-Aware Fairness Checks
Our Fairness Tooling confirms that every dataset in the Data Commons meets the standards of equity, transparency, and trust. Universal and domain-specific checks identify bias, imbalances, and coverage gaps, while automated fairness reports and mitigation workflows help rebalance data before release. Integrated with synthetic data generation, this tooling provides measurable, auditable, and transparent assessments—building confidence for both data providers and consumers and supporting regulatory compliance and responsible innovation.

Without fairness checks, synthetic datasets could reinforce blind spots. Thus, embedding fairness checks in data generation pipeline can measure and resolve imbalance and publish transparency reports to data consumers. Since data is so characteristic of the domain it is

sourced from, it is advisable to check fairness both generically and focused on the domain's specific regulations.

Universal Fairness Checks: Regardless of domain, certain fairness dimensions should apply broadly:

- **Representation Balance:** This measures whether different subgroups are present in sufficient proportions.
- **Disparate Impact Balance:** This checks if outcome variables (e.g., approvals or denials in loans) differ across groups beyond acceptable thresholds.
- **Proxy Detection:** It is useful to check if seemingly neutral fields (like zip codes) strongly predict sensitive attributes in the dataset (e.g., race).
- **Coverage and Completeness:** Are geographic areas, age ranges, or categories missing from the dataset?
- **Toxicity and Stereotypes:** Especially in text, images or speech, is there occurrence of harmful associations or language.

Domain-specific Fairness Checks: Fairness also varies by application domain and sector. The tooling must adapt checks to align with regulatory, ethical, and operational needs of the domain.

- **Finance:** We should apply the 80% rule[2] (a standard heuristic for assessing fairness) for disparate impact, evaluate statistical parity and enable equalized odds for credit outcomes.
- **Health/ Bio-medical:** Verify representation balance across patient cohorts, subgroup calibration, and absence of systematic missingness.
- **Environmental/ Climate:** Confirm geographic coverage across vulnerable regions and equitable placement of sensors.
- **Real Estate/ Housing:** Test for redlining risk by examining geographic parity and correlations between income brackets and outcomes.
- **Employment/ Education:** Detect gender or occupational stereotypes, subgroup performance parity, and inequities in applicant or student datasets.

The goal is to include these domain specific checks in the fairness checker logic to assure data consumers that the definition of fairness is aligned and confined to the sector-specific requirements.

How the Fairness Checker works: We envision the fairness checker service to run as a structured pipeline comprising of the following key components:

- **Ingest Dataset:** Accepts uploaded or generated data.
- **Run Universal Checks:** Calculates representation, parity, coverage, proxy bias, and toxicity.
- **Apply Domain-specific Checks:** Tailors metrics depending on declared domain (e.g., finance vs. healthcare).

[2] https://labor.hawaii.gov/wp-content/uploads/2012/12/WIA-MOA-Element-7-2012-Exhibit-J-Adverse-80-Rule.pdf

- **Generate Fairness Report:** Produces a dataset "nutrition label" summarizing metrics, thresholds, and risk flags.
- **Support Mitigation Loop:** Where thresholds are not met, the service triggers augmentation via the synthetic generator (e.g., adding underrepresented rows or balancing slices).

The intended goal of this pipeline is to allow for automated gating—i.e., a dataset must pass fairness thresholds or carry attached warning/disclaimer before being shareable or downloadable in the marketplace. In terms of fairness tools that can be utilized depending on the specific modalities, potential options include:

- **Tabular Data:** AIF360, Aequitas, Themis-ML, Fairlearn.
- **Text Data:** WEFE, Perspective API, Checklist
- **Audio/ Speech Data:** FairEva
- **Vision:** IBM Vision 360, Azure Face API, FairFace.
- **Multimodal:** Responsible AI Toolbox

Example Scenarios: Here we describe two potential scenarios, one closer to Release 2 and one in the future as a feature in the final product.

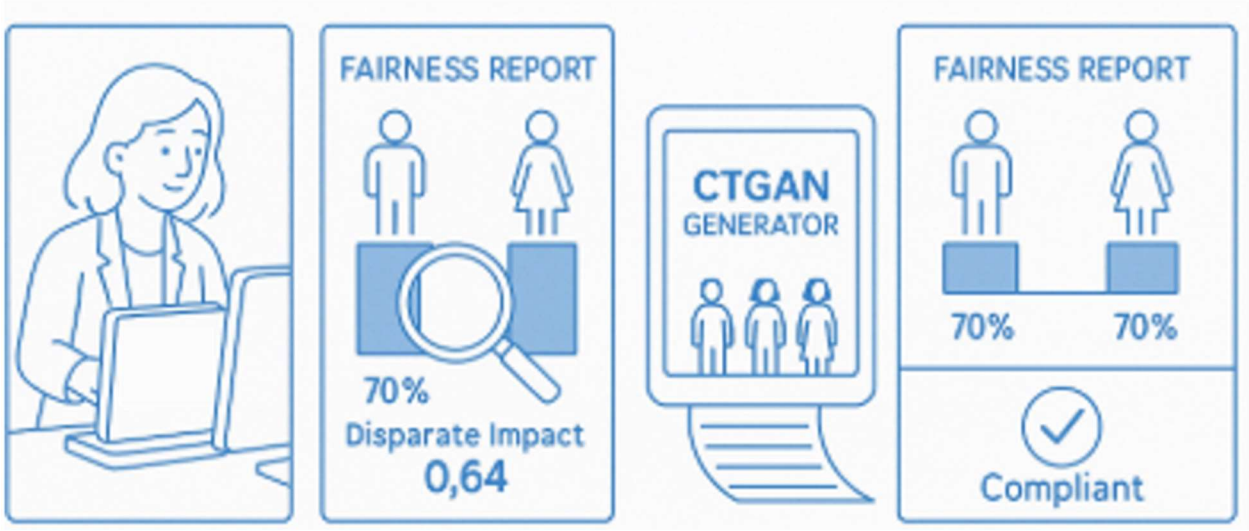


Figure 20: Example Scenarios

Scenario 1: A loan dataset is uploaded in tabular format. The checker finds female approval rates at 40% compared to 70% for males (disparate impact of 0.64). A fairness report is generated. A conditional generator (CTGAN) is then used to synthesize additional female applicants until the balance is restored, after which the checker verifies compliance.

Scenario 2: A self-driving dataset, containing autonomous driving point cloud data, is audited. The fairness checker finds only 10% of frames represent night or rainy conditions. The planner sets quotas, and a simulator (CARLA, AirSim, or NVIDIA Omniverse) generates new frames under-sampled slices (e.g., night + rain with pedestrians). The fairness tool re-audits coverage until quotas are met, then attaches a fairness card.

Benefits of Fairness Tooling: The most important benefit is trust and transparency. Data consumers see the fairness scores before using the datasets. Secondly, compliance readiness is a big factor. The tool supports adherence to regulatory standards such as fair lending laws or clinical trial diversity requirements. Thirdly, it is key for the system to provide operational assurance. This can be achieved by validating that models trained on synthetic data perform consistently across groups and scenarios. Finally, in order to achieve marketplace integrity, we prevent biased or incomplete datasets from being published unchecked. This approach creates confidence for both dataset providers and consumers, positioning the marketplace as a trustworthy and compliant environment for synthetic data innovation.

Fairness tooling is a cornerstone of responsible synthetic data services. It provides measurable, auditable, and transparent assessments of bias, enabling datasets to be rebalanced and validated before release. By combining universal checks, domain-specific modules, and allowing integration with synthetic data generators, the fairness checker confirms that synthetic data is not just statistically plausible, but also equitable, useful, and safe.

Detailed Tools: We list the potential fairness checker tools that can be integrated in the platform or utilized via API. The priority list is in decreasing order and based on the complexity of data modality and the ease of using the tool. This ranking does not consider the corpus of datasets that are chosen for initial release, which might re-rank the priority.

TOOL NAME	MODALITY	INSTALLABLE OR API	RECOMMENDED PHASE
AIF360 (IBM)	Tabular	Installable	MVP
Aequitas	Tabular	Installable	MVP
Themis-ML	Tabular	Installable	MVP
Fairlearn (Microsoft)	Tabular	Installable	MVP
Scitkit-Fairness	Text	Installable	Scale-Up

TOOL NAME	MODALITY	INSTALLABLE OR API	RECOMMENDED PHASE
WEFE	Text	Installable	Scale-Up
Checklist (MSR)	Text	Installable	Scale-Up
FairEVA	Audio/Speech	Installable	Scale-Up
FairFace	Vision	Installable	Scale-Up
Responsible AI Toolbox (Microsoft)	Multimodal	Installable	Scale-Up
Perspective API (Google)	Text	API	Scale-Up
Azure Face API	Vision	API	Final / Later
IBM Vision 360	Vision	API	Final / Later

Key Topic: Synthetic Data Generation

The Synthetic Data Generation service turns data constraints into opportunity—enabling safe, equitable, and scalable access to information for research and innovation.

Our goal is to enable the creation of realistic, high-quality datasets that preserve the statistical properties of real data while reducing risks associated with privacy, availability, and fairness. In many sectors, access to rich and representative datasets is constrained by regulation, confidentiality, or the cost of collection. Synthetic Data Generation (SDG) addresses these challenges by providing controlled, on-demand generation of data that can be used for testing, model development, analytics, or data sharing.

Our approach to SDG emphasizes two pathways:

- **Augmentation:** Starting with a seed dataset and expanding, balancing, or transforming it.
- **Organic Generation:** Creating datasets entirely from scratch, guided by schema and user specifications.

Augmentation will be the focus of early phases (Beta and MVP), where real seed datasets provide structure and quality assurance. Organic generation—creating entirely synthetic datasets—will be introduced in later phases once governance, fairness, and evaluation mechanisms are fully proven.

The earlier releases (Beta and MVP) focus on tabular data, the most common format for business applications, with extensibility to other modalities (text, image, audio, video, point cloud) in subsequent releases.

Why SDG Matters: There are many important reasons why synthetic generation tools have grown in capability and usage. Some of the key ones are:

- **Privacy Protection:** SDG helps avoid exposing sensitive information, enabling safe data sharing and compliance with privacy laws.
- **Fairness Improvement:** It can correct imbalances in real datasets by generating under-represented groups or conditions.
- **Scalability:** Provides unlimited data volume for testing or training without the cost of collecting real-world samples.
- **Coverage:** Supports adequate representation of rare events or conditions (e.g., fraud cases in finance, nighttime driving scenarios in autonomous vehicles).
- **Innovation Enablement:** Makes data available earlier in the product lifecycle, before production systems or field trials generate sufficient data.

Safe, Fair, and Scalable Synthetic Data Generation.
Our Synthetic Data service enables the creation of realistic, high-quality datasets that protect privacy, improve fairness, and support innovation. The approach starts with augmenting real data to fill gaps and reduce bias, then expands to organic generation across multiple modalities. Integrated fairness checks and privacy controls confirm that every synthetic dataset is scientifically rigorous and ethically sound, empowering users to innovate confidently while maintaining trust and compliance.

How the SDG Service Works: This section describes the front-end user interaction model and connects it to the essential system elements that power it. In alignment with our Phasing Strategy, the SDG capability evolves across releases—starting with simpler tabular augmentation in Beta and MVP, and expanding to higher-complexity, organically generated multimodal data in Scale-Up.

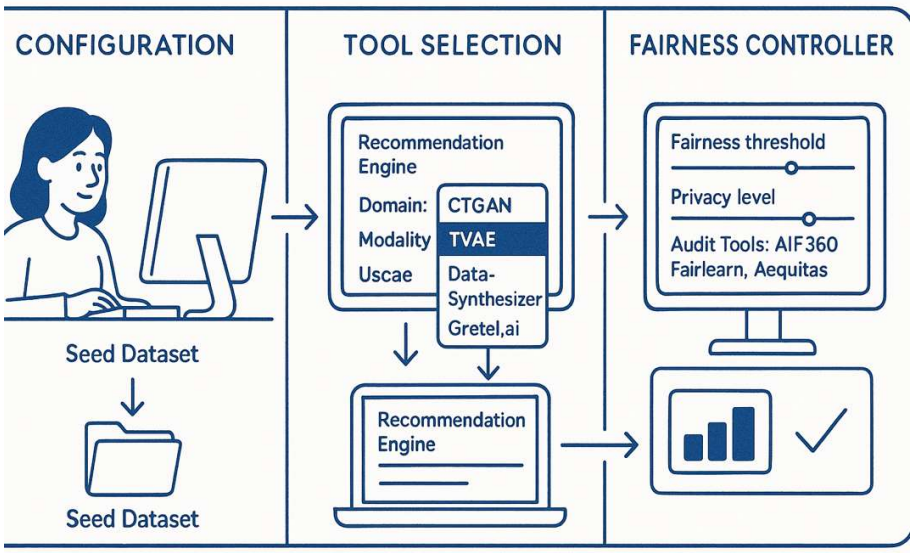


Figure 21: User-Facing Synthetic Data Generation Process

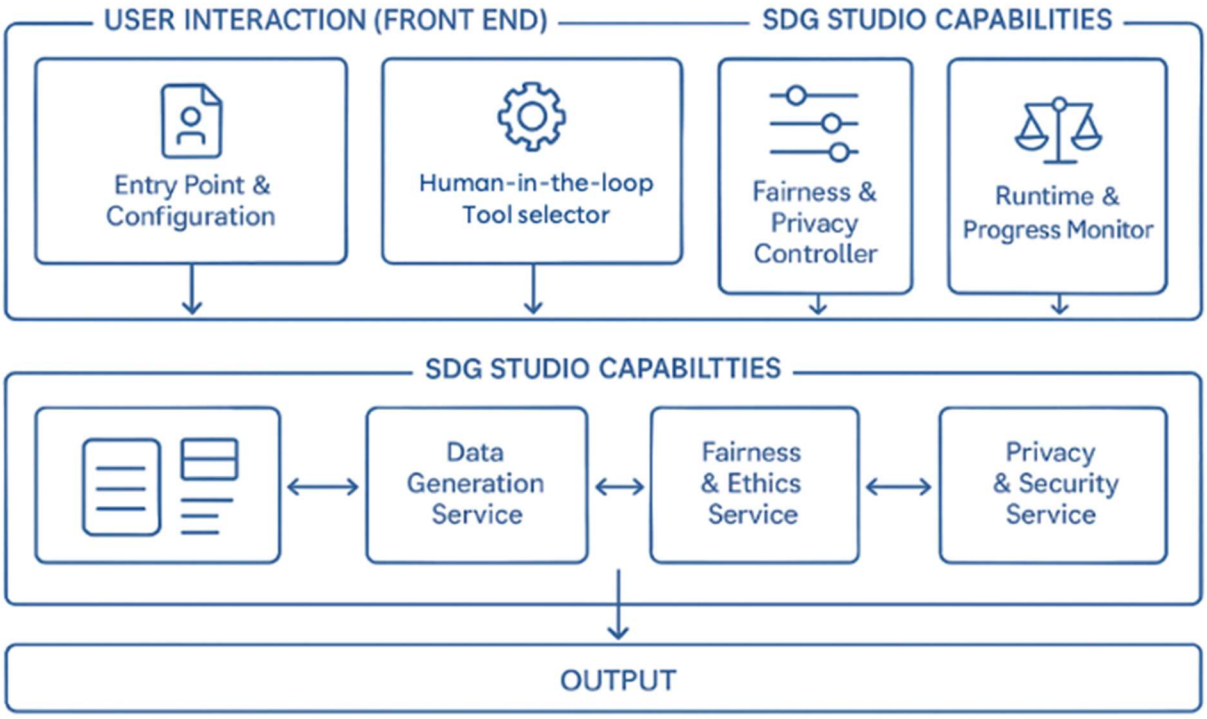


Figure 22: Synthetic Data Studio User & System Overview

From a user perspective, the first interaction with synthetic data generation is **Entry Point and Configuration**. The synthetic data generation process begins with a simple but structured entry point designed to guide the user through clear configuration choices. When initiating a generation task, the user specifies the modality of the dataset they wish to create or augment. In the first release, the supported modality is tabular data, which covers common use cases such as financial transactions, healthcare records, or retail purchase logs. Future releases will extend this framework to other modalities, including text, images, audio, video, and point clouds.

In addition to modality, the user also declares the domain and intended use case of the dataset. This information is critical because different domains often imply different fairness risks and tool preferences. For example, a financial dataset might require tabular synthesis using CTGAN (Conditional Tabular GAN, part of the SDV library) to handle mixed categorical and continuous features while also enabling fairness-aware augmentation. By contrast, a healthcare dataset may benefit from the DataSynthesizer for privacy-preserving statistical modeling, while a retail dataset might leverage something like "GaussianCopula" models from the SDV (Synthetic Data Vault) suite to capture correlation structures between features. For tabular data where interpretability and reproducibility are important, a rule-based tool like Synthpop (R) can also be selected.

Users then provide a data specification to further shape the generation process. This may include schema definitions (columns and data types), the number of rows to generate, or statistical constraints such as distributions for certain fields. Where available, users can also upload seed datasets. These seeds serve as templates, allowing the generator to model real data patterns and then expand them with synthetic records that maintain fidelity while reducing bias or protecting privacy.

Our second task is **Tool Selection**. Once the configuration is established, the system employs a recommendation engine to automatically select the most suitable synthesis tool based on the combination of modality, domain, and user goals. For tabular datasets in finance, CTGAN or TVAE (Tabular Variational Autoencoder) from the SDV ecosystem are often recommended, as they are designed to handle imbalanced categorical variables and preserve non-linear feature relationships. For privacy-sensitive contexts, tools such as DataSynthesizer can be chosen to enforce differential privacy guarantees during generation. Commercial platforms such as Gretel.ai may also be recommended when enterprise-grade scalability and compliance features are required.

In other domains outside tabular data, which are targeted for future releases, the recommendation engine would suggest different classes of tools. For text generation, large language models (LLMs) such as the GPT family, Claude, or specialized frameworks like Snorkel and NL-Augmenter could be invoked. For image synthesis, state-of-the-art generative adversarial networks like StyleGAN2/3 or diffusion models such as Stable Diffusion and Imagen would be appropriate, particularly when paired with fairness-aware datasets like FairFace to rebalance demographic representation. In simulation-heavy contexts like autonomous driving, the engine could recommend CARLA, AirSim, or NVIDIA Omniverse Replicator to generate point clouds and labeled 3D environments. This recommendation layer removes the burden from end-users of knowing which algorithm to select. The system matches configuration inputs with appropriate tools, so each dataset is created using methods that fit its technical needs and fairness goals. Our recommendation would be to start with tabular generation, and then move into text, image and finally video. This is because not only is it computationally cheaper to produce tabular and text but also fairness tools in this space are easier to integrate than the latter.

The third step is a **Fairness Controller**. After the recommendation step, users can tailor the dataset further by selecting fairness and privacy preferences. The platform exposes these preferences through simple controls such as sliders, toggles, or policy files, making complex fairness and privacy trade-offs accessible even to non-experts. For fairness, users may set thresholds for acceptable subgroup disparity or require balanced representation across sensitive attributes such as gender, age, or geography. For example, in a loan approval dataset, the user could mandate that approval rates between male and female applicants remain within 80% of each other, aligning with the “four-fifths” rule commonly used in regulatory contexts. These fairness preferences integrate directly with the fairness checker tools (e.g., AIF360, Aequitas, Fairlearn) that audit both the original and synthetic datasets. If the system identifies imbalances during audit, it can automatically trigger conditional generation modes in tools like CTGAN or Gretel.ai to synthesize under-represented groups until the fairness thresholds are satisfied.

With modality, domain, tools, and fairness/privacy settings defined, the system proceeds to the runtime stage. Here, the platform focuses on user feedback and transparency throughout the generation process. Before starting synthesis, the system provides an estimate of runtime and resource costs, giving users clarity on what to expect. Once generation begins, a progress bar and log of actions (training GANs, sampling rows, applying privacy noise, etc.) keep users informed.

When the process is completed, the platform produces both the synthetic dataset and a Dataset Card (or "nutrition label"). This card is a structured metadata report that documents the dataset's provenance, the toolchain used (e.g., CTGAN v1.7, trained on seed data with 100k rows), fairness and privacy scores (e.g., disparate impact = 0.92, differential privacy epsilon = 1.0), and utility comparisons against the original dataset (e.g., correlation preservation within ±5%). The Dataset Card is automatically attached to any dataset published into the marketplace, so downstream consumers can evaluate its trustworthiness.

Thus, there are four real capabilities of the SDG studio:

1. **Data Generation –** Maintains modality specific generators, supports augmentation and from-scratch generation (future) and optimizes for cost and scalability.
2. **Fairness & Ethics –** Defines the universal fairness checks and integrates the fairness checker service.
3. **Data Quality –** Generates utility reports and scorecards about fidelity and diversity.
4. **Privacy & Security –** Can generate data under privacy-preserving synthesis such as differential privacy GANs. It can also transform or redact sensitive attributes to prevent identification.

Example Scenarios: Here are two example scenarios (earlier releases and final production version) demonstrates what the user journey could look like:

Scenario 1: In early releases, the synthetic data generation service will focus on tabular datasets, which are the most widely used across industries such as finance, healthcare, and retail. Consider a financial institution that uploads a dataset of 50,000 loan applications. When the dataset enters the pipeline, the fairness checker (e.g., AIF360 or Aequitas) evaluates approval rates across sensitive groups such as gender. In this case, the audit reveals that approval rates are 70% for men and 40% for women, resulting in a disparate impact ratio of 0.64, which falls below the regulatory threshold of 0.8 (i.e., the 80% rule). The system then applies a default fairness policy to restore balance by invoking a tabular generator such as CTGAN. Using conditional generation, the system synthesizes additional records for underrepresented female applicants so that the new rows preserve statistical correlations with income and credit score. After merging the synthetic and original records, the fairness checker re-runs its analysis and confirms that the dataset now meets parity requirements. At this stage, the service produces a Dataset Card summarizing fairness and utility scores and gives the user options to download the balanced dataset, publish it to the marketplace, or revert to the original version.

Scenario 2: In a final production scenario, the service will extend to more complex modalities such as point cloud data for autonomous driving. Suppose a dataset of LIDAR frames is uploaded where most scenes capture clear, daytime urban driving. The fairness audit shows

significant underrepresentation of adverse weather conditions, nighttime driving, and vulnerable road users such as cyclists or pedestrians at distance ranges. To address this imbalance, the system computes a coverage matrix and sets quotas for additional slices (e.g., +8,000 frames of night + rain conditions, +1,000 scenes with cyclists at mid-range distances). A simulation-based generator such as CARLA, AirSim, or NVIDIA Omniverse Replicator is then invoked to produce synthetic frames under the missing conditions, with automatic labels generated from the simulator (bounding boxes, tracks, semantic segmentation). The dataset is re-audited until the fairness thresholds are met and coverage across environments, lighting conditions, and actor types is balanced. The final version is frozen, documented with a Dataset Card that reports fairness, coverage, and utility metrics, and made available for download or publication in the marketplace.

Detailed Tools: Before we detail specific tools, it's helpful to explain how they will connect to the platform over time. There are three levels of integration, each representing a step in the platform's maturity. In keeping with our smart sequencing strategy, we begin simple and deepen integration as we learn:

1. **External:** The platform recommends the tool and routes the user to its external landing page to run it independently. (Fastest to implement; used in Beta.)
2. **Semi-integrated:** The tool is hosted within the platform, but configuration occurs through a linked interface. (Introduced in MVP.)
3. **Fully integrated:** The tool runs natively in the platform, preconfigured by the recommendation engine with minimal user input. (Targeted for Scale-Up.)

In early releases, the recommendation engine will primarily point to external or semi-integrated tools—such as routing to FairFace or Gretel.ai with pre-populated configuration details. Over time, the system will support direct in-platform execution, creating a seamless experience without compromising early delivery speed or flexibility.

The priority list is in decreasing order and based on the complexity of data modality and the ease of using the tool. This ranking does not consider the corpus of datasets that are chosen for initial release, which might re-rank the priority.

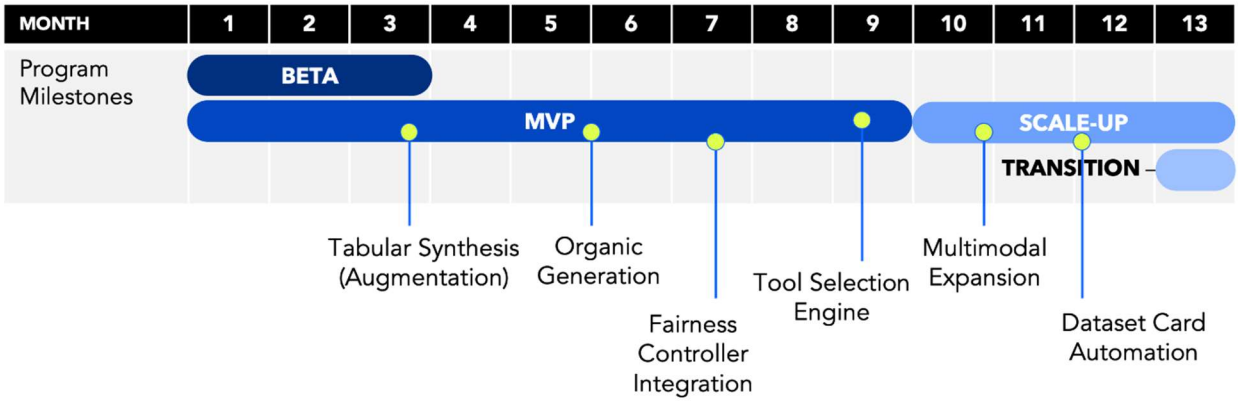


Figure 23: Illustrative Synthetic Data Generation Feature Timeline

TOOL NAME	MODALITY	INSTALLABLE OR API	RECOMMENDED PHASE
SDV (CTGAN, TableGAN)	Tabular	Installable	Beta
DataSynthesizer	Tabular	Installable	MVP
Synthpop (R)	Tabular	Installable	MVP
Gretel.ai	Tabular/ Text	API	MVP
Hugging Face Toolkit	Text	Installable	Scale-Up
Snorkel	Text	Installable	Scale-Up
Parrot (MSR)	Text	Installable	Scale-Up
DataPerf	Text	Installable	Scale-Up
GPT-family	Text	API	Scale-Up
AWS Polly	Audio/Speech	API	Scale-Up
StyleGAN2/3 (Nvidia)	Image	Installable	Scale-Up
ProGAN	Image	Installable	Scale-Up
FastSpeech	Audio/Speech	Installable	Scale-Up
AutoVC	Audio/Speech	Installable	Scale-Up
SpecAugment	Audio/Speech	Installable	Scale-Up
Imagen (Google)	Image	API	Final / Later
CARLA	Point Cloud/ Video	Installable/Simulator	Final / Later
AirSim (Microsoft)	Point Cloud/ Video	Installable/Simulator	Final / Later
Nvidia Omniverse	Point Cloud	Installable/Simulator	Final / Later

Together with the Fairness Checker, the Synthetic Data Studio enables the DCC to provide synthetic datasets that are both scientifically rigorous and ethically sound—advancing innovation while protecting the public trust.

Key Topic: User-Contributed Data

User-Contributed Data

While curated datasets establish initial credibility, long-term value depends on enabling and incentivizing community contributions.

We will make contributing data straightforward and rewarding. Contributors will be able to upload a dataset through a clear, step-by-step workflow, see it published once approved, and receive credit when others use it. Visibility and acknowledgment are built into the process. Commons Credits will serve as an additional incentive, but the primary driver is a well-designed, transparent user experience.

Simple, Rewarding, and Trustworthy Data Contribution
Our approach makes contributing data to the Commons straightforward and motivating. Guided workflows and clear governance support quality and transparency, while phased release strategies expand features and incentives over time. Contributors are recognized for their impact, earning Commons Credits and seeing their datasets used by others. This balanced model builds trust, encourages participation, and sustains a vibrant, high-quality data ecosystem.

Contribution Challenges and Our Approach

Different contributors bring different expectations. Healthcare researchers may be eager to share, given their experience with other data commons. Startups may be cautious, concerned about competition. Civic innovators may be motivated by impact but lose patience with slow or complex processes. These differences make contribution more complex than it first appears.

We address this with a human-centered approach and smart phasing:

- **Workflows and UI** make contribution simple and satisfying, with guided forms for classification, sensitivity, usage restrictions, and provenance. Contributors can set key attributes up front rather than navigating complicated process.
- **Governance** maintains quality without unnecessary friction. Metadata is validated automatically, curators review submissions, and required fields like license and sensitivity are enforced.
- **Phased Release Strategies** help keep initial development scope realistic while expanding utility over time. Initial submissions focus on essential FAIR attributes. Later phases introduce more advanced features like dataset versioning, FAIRness scoring, and automated enrichment.

Phased Approach to Contributions

A phased strategy accelerates learning about end-user behaviors and adds value as adoption grows.

- **Phase 0 (Beta):** We suggest starting with a simple manual submission form, reviewed by a staff member. Consent defaults to "full open use." Metadata requirements are minimal but mandatory. Workflows will guide contributors step-by-step to make the process manageable. User-contributed datasets will be flagged with a simple visual indicator.
- **Phase 1 (Public Release):** Automated validation will be layered in to reduce review time. Credit rewards could start as a flat reward. A limited license/consent menu would be introduced. Expansion of contributions beyond data could include blog posts, instructions, and documentation—all credit-eligible.
- **Phase 2 and Beyond:** Expansion to more data modalities (e.g., images, point clouds) and sensitive domains, along with links to datasets hosted elsewhere. Credits become more dynamic, with recurring rewards for high-use datasets. User reviews and feedback introduced, with moderation and credit rewards for reviewers.

Governance and Quality

Quality and governance are essential to building trust. From the first release, all contributed datasets will be required to include basic metadata and provenance information. These minimum standards make datasets findable, reusable, and citable.

To manage risk and adoption, early phases will focus on data types and domains with simpler standards. This defers the friction that can arise from complex quality checks until the model is more fully proven while establishing clear expectations. Over time, more advanced validations like FAIRness scoring, automated metadata enrichment, and lineage capture will be layered in.

For contributors, this process is designed to feel straightforward: guided forms help set classification, sensitivity, usage restrictions, and provenance. Automated checks catch missing fields or errors. Curator review maintains quality and consistency without unnecessarily slowing down the workflow. This balance preserves efficiency while upholding baselines for quality and trust across every published dataset, whether community submitted or provided by the DCC.

Consent and Licensing

During the Beta phase, user-contributed datasets will default to open-use licensing to reduce friction and accelerate adoption. Beginning with Release 1, contributors will be able to choose from a limited menu of license and consent options covering common open, academic, and restricted-use cases. Contributors may propose terms, but final approval rests with governance staff to confirm consistency and compliance. Over time, the menu of options will expand to cover additional domains and sensitivity levels guided by community input and governance policy.

Motivation and Rewards

The primary motivation for contributors will come from the experience itself. A well-designed workflow makes the act of publishing data straightforward and satisfying. Contributors see their datasets approved and listed in the catalog with clear visibility into how often they are accessed or reused. That feedback loop—knowing that a contribution is visible, trusted, and useful to others—is a form of recognition that reinforces intrinsic motivation.

Alongside that intrinsic value, the system will offer extrinsic rewards through Commons Credits. In the early phases, credits will be granted as flat rewards for accepted contributions to keep the model simple. Over time, this will evolve into more dynamic incentives like recurring credits for datasets that see high levels of use. Credits will not only recognize contributions of new data, but also the important work of reviewing, curating, and validating datasets submitted by others. This balanced approach sustains engagement while actively reinforcing quality and governance.

Building Trust Over Time

In Beta, trust will come from simplicity and oversight. Every contribution will go through staff review, and datasets contributed by users will be clearly marked in the catalog. This means that early adopters can see what is user-submitted while knowing it has passed a baseline quality check.

With the first public release, additional safeguards will be introduced. Automated checks will verify metadata completeness, sensitivity flags, and provenance, while a small menu of license options gives contributors flexibility without overwhelming the process. Contributors will also be able to add supporting content like documentation, usage notes, and guidelines that make their datasets easier for others to understand and reuse.

In later releases, trust will be reinforced by breadth and community participation. More data types will be supported, Commons Credits will move beyond flat rewards to reflect usage and impact, and community reviews will provide another layer of validation. Over time, contributions will extend beyond datasets to include models, tools, and annotations—broadening the value of the Commons while keeping quality and governance at the center.

Solution Conclusion

Taken together, these components form a complete solution for the Data Commons—balancing vision with practicality to deliver measurable impact early while laying the foundation for long-term success.

Appendix - Team & Staffing

Within this section, we include the key personnel, and their qualifications, of the individuals intended to perform the work. Depending on the exact timing of the project start, it is possible that the proposed team members may be replaced with similarly experienced and

qualified resources. Slalom intends to align key subject matter experts to this project to maintain technical expertise and knowledge of the proposed platform.

Barry Scharfman, Ph.D.

Head of Data & Intelligence Capability - Boston, MA – Subject Matter Expert

15+ years in Data Science and Consulting driving value creation through transformation and delivery of insights and tools. Functional experience includes AI / ML, Data Analytics and Visualization, BI, Agile transformation, and strategic and operational performance improvement across industries including Industrials, Manufacturing, Aerospace & Defense, Pharma, Public Sector, and Private Equity. M.S. and Ph.D. from MIT in Engineering, dual B.S. in Engineering and Business (Wharton) from the University of Pennsylvania. U.S. Dept. of Defense Top Secret Security Clearance.

RELEVANT EXPERIENCE

GenAI Tool Development Private Equity
For large cap and mid-market Private Equity clients, developed GenAI tools to identify investment targets and recommend performance improvement initiatives for target and current portfolio companies.

Production Simulation Modeling Industrials Manufacturing
As part of a Private Equity buy-side due diligence on a target manufacturer, used Monte Carlo simulation to model impact of various manufacturing, supply chain, and logistics issues on production rate of precision engineered platform.

AI / ML / Computer Vision Public Sector
Led a team to develop an AI / ML Computer Vision platform funded by a federal government client for global private and public sector safety and security applications.

Manufacturing and Supply Chain Aerospace & Defense
Led team to develop complex discrete event simulation model and data visualization for manufacturing, maintenance, and upgrades schedule of major Defense aircraft platform.

Product Owner Public Sector
Served as Product Owner for newly introduced public-facing digital communications platform for federal client. Improved customer adoption (>1K users during pilot phase) and satisfaction (>80%), increased efficiency (decreased case cycle time by weeks), and reduced costs (≈70% call deflection).

SKILLS

- Data Science / GenAI / Machine Learning
- Data & AI Strategy
- Agile Transformation & Product Ownership
- Design Thinking
- Business Intelligence & Analytics
- Programming Languages: Python, R, etc.
- Data Analysis & Visualization: Tableau, Power BI, Alteryx, ArcGIS, etc.
- Operational Performance Improvement
- Government Contracting
- Due Diligence

Chris Patten

Director, Data Architecture & Engineering – Subject Matter Expert

Chris is a seasoned data leader with a strong technical background across many industries and technology platforms. He excels at bringing together business and technical leaders to drive to outcomes.

RELEVANT EXPERIENCE

Platform Architect – Healthcare Risk Management
Chris led a team of architects, engineers, and analysts to completely modernize the data management and analytics capabilities of a 30+ year old captive medical professional liability insurer. Chris architected a new self-service analytics platform using Snowflake, Fivetran, dbt, and Power BI, and directed his team in implementation of the future-proof, cloud native solution that enabled best practices in DevOps and CI/CD for data warehousing.

Data Architect – Marketing Automation Platform
Chris designed and led development of a Snowflake-based lead scoring solution for a large healthcare company. The scoring model served as the lynchpin of a large reimagining of the lead management process and combined data from Salesforce, Eloqua, and other systems in a user-configurable model designed from the ground up to support analysis and tuning.

Data Architect – Passenger Railroad
Chris led the legacy system technical analysis and modernization planning for a custom operations platform that blends live data from multiple sensors and onboard systems for a major commuter rail operator.

Data Architect/Engineer – Enterprise Software M&A
Chris led a large cross-functional team integrating data as part of a major acquisition for an enterprise software company. Chris and team reconciled complex business rules from both organizations and developed data ingestion processes to merge data across CRM, CPQ, and customer support domains.

SKILLS

- Program leadership
- Use case and value case development
- Data strategy
- Data governance and enablement
- Senior Snowflake architect
- Modern data architectures
- dbt
- AWS, Azure
- Power BI architect
- SQL, Python
- Data product management
- Building teams and leading with empathy

Roderick (Rick) McMullen

Director, Business Transformation – Delivery Solution Lead, Subject Matter Expert

Rick has 20 years of experience leading end-to-end customer-driven innovation—from concept through market impact and solution delivery. He has served as a founder/CEO, P&L leader, product executive, and award-winning UX designer. Rick holds an MBA from Harvard University and dual degrees in Human-Computer Interaction from Carnegie Mellon University.

RELEVANT EXPERIENCE

Founder & CEO – Health & Wellness Venture
Founded and grew a health and wellness start-up that raised $6M in funding and achieved $3M in revenue. Launched four product lines for chronic injury treatment and earned a Red Dot Award for medical design. Developed business model and go-to-market strategy; now serves as Board Director.

General Manager, Supply Chain & Operations – Wayfair
Managed a $30M P&L, leading a tech and operations transformation that increased throughput 40% and cut costs 25%. Drove Agile adoption, launched products improving customer conversion by 9%, and led warehouse automation initiatives boosting delivery speed and efficiency.

Designer & Innovation Coach – Global Design Consultancy
Designed award-winning products and led hundreds of Human-Centered Design workshops across industries including healthcare, finance, defense, and logistics. Specialized in simplifying complex, information-rich systems for user and business impact.

SKILLS

- Business Model Design & Change
- Customer Loyalty Strategy & Design
- Product Strategy & Product Mgmt
- Human-Centered & Experience Design
- Supply Chain & Operations Mgmt
- Innovation Strategy & Process
- Storytelling & Vision
- Executing Innovation in Organizations

Dennis Rizzi

Senior Director, Solution Ownership – Accountable Executive

Dennis has 25+ years of IT experience in PSI and commercial industries, product management and program management with a focus in Agile Delivery. He has led complex, large programs across various IT disciplines including software development, data transformations, cybersecurity and cloud infrastructure.

RELEVANT EXPERIENCE

Product Delivery Lead – Data Solutions (Public Sector)
Oversaw delivery, schedule, and quality across multiple public sector data transformation, Snowflake, and AI/ML solution engagements.

Senior Solution Owner – Enterprise Services Modernization (Federal)
Led planning, delivery, and maintenance of seven integrated systems for the Department of Homeland Security using Enterprise Agile and SRE practices.

Senior Solution Owner – Enterprise Services Modernization (Financial Services)
Directed discovery and architecture for re-platforming a trading system into a modular, microservices-based platform. Delivered a six-month proof of concept integrating enterprise services.

Solution Owner – Data Transformation (Telecom)
Redesigned and migrated a multi-service capability into an AWS-based data transformation architecture, improving customer insights and advertising impact.

Product Manager – Data Management (Healthcare)
Managed product strategy and roadmap for six products across healthcare, financial, and government markets. Defined data governance policies, conducted analysis, and supported go-to-market planning.

SKILLS

- Project/Program Management
- Product Management
- Agile Software Development
- Data Analytics
- Modern Web Development
- Enterprise Program Delivery
- Project Management Professional (PMP)
- Certified Scrum Master (CSM)

Colleen DaRos

Delivery Director, Delivery Leadership – Program Manager

Colleen has 20 years of consulting experience focused on effective project delivery. She is passionate about leading high-performing teams with a focus on outcomes, and agile/scrum delivery at the enterprise level.

RELEVANT EXPERIENCE

Engagement Lead – State of Massachusetts, Salesforce Case Management
Directed discovery, design, and development of a Salesforce Case and Document Management System to replace outdated technology, serving multiple courts and agencies with a public-facing portal and integrated external data exchange.

Engagement Lead – State of Massachusetts, Salesforce Grants Management
Led implementation of Salesforce Grants Management Solution supporting applications, awards, and reimbursements for 351 municipalities. Delivered public portals, municipal interfaces, and reviewer tools for grant scoring and award management.

Engagement Lead – Healthcare, Digital Patient Experience Program
Led post-launch CRM and DevOps enhancements following a major digital front door transformation. Strengthened platform capabilities and processes to improve patient engagement and streamline marketing, web, and call center operations.

SKILLS

- Project / Program Management
- Mobile / Responsive Web Design
- Functional Analysis
- Salesforce CRM
- Content Management Systems (CMS)
- Agile/Scrum Delivery
- Customer Engagement

Laura Kenney

Delivery Leadership – Chief Product Owner

Laura is a consulting professional with 13 years of experience leading cross-industry projects in financial services, higher education, life sciences, and the public sector. She excels at identifying organizational challenges and delivering impactful solutions as a business architect, product owner, project manager, and business analyst.

RELEVANT EXPERIENCE

Business Architect – Retail
Facilitated workshops to assess strategic drivers, current-state architecture, and design a future-state Sales and Service model to merge two organizations into a unified global operation.

Business Architect – Public Sector
Led discovery and planning workshops to define business needs and solutions, ensuring alignment across technology, processes, and end-user requirements as part of a leadership team.

Engagement Lead / Business Architect – Healthcare & Life Sciences
Directed a multi-year transformation to expand a center of excellence, conducting stakeholder workshops and developing a roadmap to improve efficiency, quality, and cost savings.

Product Owner & Project Manager – Professional Services
Engaged 40+ stakeholders to identify taxonomy gaps and process challenges, facilitating workshops and developing a roadmap to design and implement enterprise knowledge management enhancements.

Engagement Lead & Business Strategist – Financial Services
Led development of a predictive attrition model, aligning business and technical teams to deliver insights that improved customer retention and lifecycle management.

SKILLS

- Process Analysis & Improvement
- Salesforce (Commerce, Sales, & Service Cloud)
- Project Management
- Product Ownership
- Communications & Stakeholder Engagement
- Change Management
- Vendor Selection
- Center of Excellence / Operating Model
- Workshop Facilitation

CERTIFICATIONS

- Prosci Change Management Practioner

Jen Feldman

Customer Strategy, Experience Strategy & Design – Experience Strategist

Jen is a seasoned experience designer who solves problems with a user-centered approach. She is highly skilled at turning complex processes into simple, seamless, and intuitive experiences for users. Jen excels at promoting customer needs while aligning her designs to business goals and objectives.

RELEVANT EXPERIENCE

Market Intelligence Platform
Led workshops to define AI use cases and designed user flows and prototypes, including an AI-driven insights page with chat-based Q&A for intuitive data interaction.

Data Management Solution
Defined product strategy and design for a financial data platform through research, wireframing, and prototyping, streamlining workflows and modernizing the user experience.

Website Modernization
Directed a design team in modernizing state agency websites. Conducted research, developed new navigation and content structures, and created wireframes and mockups.

COVID Vaccine Portal
Designed a statewide vaccination portal, conducting rapid research, prototyping, and stakeholder presentations to guide iterative improvements.

State Agency Benefits Portal
Contributed as UX designer on a high-profile public sector project, developing personas, user journeys, prototypes, and conducting usability testing.

Life Sciences Organization
Redesigned an e-commerce platform to boost engagement and sales through user research, high-fidelity prototyping, and usability testing.

SKILLS

- Design Thinking
- Personas
- Journey Mapping
- Wireframing
- Interactive Prototyping
- User Interface design
- User Research & Usability Testing
- Design Systems

TOOLS

- Figma/FigJam
- Sketch
- Adobe Creative Suite
- Miro and Mural
- Userlytics

CERTIFICATIONS

- UX Mini-Masters from Rutgers University

Courtney Silva

Director Transformation – Enablement SME

Courtney is a seasoned Change Management professional with 15+ years of experience advising, leading, and managing the success of programs focused on guiding behavior changes to enable business transformations. Courtney has honed her skills in Change Management, including building enterprise-wide change capabilities; driving technology adoption efforts for global, large-scale transformations; and guiding leadership alignment across a variety of industries and geographies.

RELEVANT EXPERIENCE

Change Advisor – Public Sector (Local Public Health)
Led a 7-person change team within a $35M program driving adoption of a statewide public health data platform across 351 municipalities. Developed culturally relevant training strategies and led cross-functional alignment to improve data use and health outcomes.

Change Advisor – Life Sciences
Guided a 4-person team through a large-scale ERP implementation, facilitating leadership alignment and managing change networks, communications, and engagement events to accelerate adoption and business value.

Change Advisor – Healthcare
Directed global change management for a Salesforce-based customer experience transformation. Partnered with leadership across six workstreams to deliver communications, training, and adoption strategies for a 70-person global team.

SKILLS

- Change Management
- Change Agility
- Building Change Capabilities
- Enterprise Change Offices
- Learning Design & Development
- Leadership Alignment
- Leadership Development

Anna Darling

Project Delivery – Client Partner

10+ years of project management experience leading cross-capability and global teams to meet the complex needs of her clients. Lead a Commonwealth of Massachusetts initiative involving multi-stream coordination, risk mitigation, and stakeholder alignment for over 1,800 users. Demonstrated ability to translate complex client problems into tangible pieces of work to maintain stakeholder alignment

RELEVANT EXPERIENCE

Project Manager – Public Sector, Commonwealth of Massachusetts
Managed three workstreams with 20+ team members to reduce recidivism by transitioning from a docket-based to an individual-based system. Oversaw Salesforce Public Sector Solutions (PSS) implementation, third-party integrations, and adoption for 1,800+ users, leading prioritization, risk mitigation, and issue resolution.

Engagement Lead – Healthcare
Led a five-person cross-capability team to automate Customer Price Index (CPI) rate adjustments using Salesforce Billing and CPQ. Partnered with the CIO and Board of Directors to deliver a $20M annual revenue increase through streamlined processes and automation.

Engagement Lead – Technology
Served as Program Manager for a 17+ member team on an 18-month, $400M acquisition integration involving data migration and Salesforce integration. Directed risk and scope management across 14 workstreams and collaborated with senior leadership to mitigate program-level risks and provide successful delivery.

SKILLS

- Project management
- Agile/Waterfall Delivery
- Risk mitigation
- Scope management
- Stakeholder management
- Milestone based planning
- Organization and time management
- Issue prioritization
- Data-based status tracking
- Collaboration and teamwork, conflict resolution
- Executive-level relationships
- Jira, Confluence, MS Office

Anisha Jain

Data Engineer

Anisha has diverse experience working in industries ranging from business services to information technology. She graduated with a Bachelor of Science degree from Tufts University where she majored in Applied Mathematics and minored in Computer Science. She is proficient in data modeling techniques, data visualization, and understanding different technology environments. Anisha is extremely passionate and looks to drive success for clients.

RELEVANT EXPERIENCE

Data Engineer – Financial Services
Modernized on-prem architecture to Snowflake using dbt, refactoring legacy code and building transformation models for analytics. Led dbt training sessions and supported data modeling and requirements analysis.

Data Visualization Developer – Healthcare
Enhanced COVID and Monkeypox dashboards for a major healthcare system, improving data visibility and insights for leadership.

Lead Business Analyst – Information Technology & Services
Created spend and headcount visualizations and financial models during a corporate demerger, presenting strategic insights to senior leadership.

Data Analyst – Computer & Network Security
Analyzed LMS data to evaluate ROI and customer retention. Configured client environments to streamline internal training processes.

SKILLS

- Business Strategy & Intelligence
- Data Visualization (Tableau, PowerBI)
- Visual Design & Best Practices
- Data Transformation (DBT)
- SQL

CERTIFICATIONS

- Tableau Desktop Specialist
- Microsoft Azure Data Fundamentals
- Agile Explorer
- Data Science Bootcamp (*Concentration: Business Analytics & Visualization)*

Suman Deb Roy

Machine Learning Engineer

A results-driven AI and machine learning leader with over a decade of experience in engineering, solution architecture and building data-driven systems. With a PhD in Computer Science, I specialize in designing and scaling ML-powered ecosystems and integrating GenAI-driven features into software platforms.

Expertise in delivering complex, LLM-driven products with deep specialization in prompting and fine-tuning Anthropic Claude, VertexAI/ Gemini and GPT families.

RELEVANT EXPERIENCE

GenAI Marketing Platform: Built an end-to-end content creation, automated copy generation and dynamic audience targeting tool with style transfer, compliance guardrails and LLM automated evaluation. Used by thousands of marketeers on AWS and content production workflow for a $27B cloud company.

Modernizing AI Infrastructure: Led the architecture redesign, migration route and model optimization for a Fortune 500 home security company, moving their ML and data systems from Google Cloud into Azure. Improved model prediction performance in real-time dashboard used by hundreds of field professionals.

SKILLS

- Automation
- Generative AI
- Data Science / ML
- Predictive Analytics
- AI Strategy
- Data Engineering
- Responsible AI
- MLOps/ LLMOps
- Computer Vision
- Data Visualization
- Robotics
- Network Science
- Cloud Infrastructure
- AI Thought Leadership
- Claude/ VertexAI

John Clarkin

Solution Architect

John is an AWS Cloud Architect and Engineer specializing in Infrastructure as Code and automated deployment pipelines. He is passionate about cloud technology and actively supports client and consultant development through AWS certification study groups and events.

RELEVANT EXPERIENCE

Cloud Migration Engineer – Media & Entertainment
Developed Terraform, GitHub Actions, and Python-based automation and patching solutions for AWS.

Cloud Engineer (IaC SME) – Insurance Payments
Led Infrastructure as Code development for AWS serverless migration. Built reusable Terraform modules with Lambda, Step Functions, API Gateway, and CloudFront. Upskilled 12+ client developers to enable scalable delivery.

Cloud Architect / Engineer – Insurance Enterprise Architecture
Architected reusable AWS Terraform modules and reference implementations to accelerate large-scale migration initiatives.

DevOps Engineer – Insurance Enterprise Architecture
Enhanced Jenkins pipelines to deploy AWS resources across multiple accounts and regions, enabling parallel execution and continuous integration.

Cloud Architect – Insurance Line of Business
Led a two-year Terraform migration of 200+ EC2 instances, developing standards, documentation, and leading seven successful migration waves.

SKILLS

- Terraform
- CloudFormation
- AWS
- GitHub Actions
- Jenkins

CERTIFICATIONS

- AWS Certified Solutions Architect - Associate
- AWS Certified Developer – Associate
- AWS Certified SysOps Administrator - Associate
- AWS AI Practitioner

Steve Small

DevOps Lead

Steve is an experienced leader in modern software delivery on hyperscale platforms. Background in operations and quality, he focuses on process and tool improvement to unlock the potential of teams. A strong communicator and mentor, Steve focuses on people, enabling first-class developer experiences that accelerate human ideas into production systems.

RELEVANT EXPERIENCE

Architect – Azure – Music & Entertainment
Platform Architecture Leader working with a comprehensive Azure-based microservices ecosystem spanning 15+ services with Node.js, .NET, and GraphQL aggregation layers, deployed via Azure Functions and AKS. Architected scalable Infrastructure as Code using Terraform with Atmos orchestration and CI/CD pipelines that streamlined enterprise deployments across multiple environments.

Architect - AWS - Cybersecurity
Led mobilization and modernization planning for a SaaS provider of malware and antivirus solutions for Salesforce. Defined architecture and proofs of concept for AWS-based modernization to improve scalability and reliability for a system achieving thousands of file scans per second.

Architect - Azure - Automotive Services
Architect for a discovery and planning engagement, defining a data engineering solution and future-state web application architecture using Azure Data Factory, Databricks, and Azure Function Apps

Architect - Azure - Automotive Retail
Worked as program-level platform leader and managed a team focused on developer enablement for a serverless Azure application with a web and mobile front-end, and API backend. Focused on site reliability and operational resilience, delivering robust incident response protocols and infrastructure as code.

SKILLS

- Hyperscale cloud platforms
- Infrastructure-As-Code
- CI/CD
- Site Reliability Engineering
- Incident Response
- Amazon Web Services
- Microsoft Azure
- DevOps Transformation

Allan Tauro

Cloud Engineer

Allan stands out as a highly adept tech consultant within Slalom's Application and Development sector, featuring an extensive background in Cloud, DevOps, architecture, and strategic security initiatives. His profound expertise makes him an indispensable contributor to a wide array of projects, from pioneering startups to Fortune 500 enterprises, with a special focus on enhancing cybersecurity measures.

RELEVANT EXPERIENCE

Omnicell – DevOps Engineer
Streamlined Azure-to-AWS migrations with Terraform, improving MAP program efficiency. Managed AWS Transfer Family, AFT customizations, and SRE initiatives; implemented Datadog monitoring and automated CI/CD pipelines across AWS accounts.

Sarepta Therapeutics – DevOps Engineer
Built a CI/CD pipeline integrating AWS, GitLab, Terraform, and Snowflake. Automated Snowflake DDL changes and produced detailed documentation for seamless handoff.

DTCC – DevSecOps Engineer
Configured HashiCorp Vault and managed secure CI/CD pipelines with Terraform. Oversaw AWS IAM roles, policies, and permissions, strengthening security and compliance.

Symbotic – Cloud Engineer
Optimized Azure environment for bot simulation using Azure Migrate and Azure Monitor. Implemented Azure AD and SSO to enhance authentication and secure user access.

Argenx – DevOps Engineer
Developed serverless APIs with Node.js and AWS services including Lambda, DynamoDB, and API Gateway. Built automated CI/CD pipelines with AWS CodeBuild and GitHub for scalable deployments.

SKILLS

- Azure Certified
- AWS Certified
- Terraform Certified
- IAM Engineer
- JAMF Certified
- DevOps Expertise
- Cloud architecture
- Systems Administration
- Security Engineer
- Software development

Luke Joliat

Release Manager – Product Enablement

Luke Joliat is a full stack developer with over five years of experience across front-end, back-end, and cloud development. An AWS Certified Cloud Practitioner with expertise in AWS and Google Cloud, he combines strong technical skills and communication to deliver effective, client-focused solutions.

RELEVANT EXPERIENCE

Front End Architect – International Consulting Services

Responsibilities included leading frontend team in architecting and creating a Shared Component system that allowed teams within the business to share artifacts. Designed and developed system using ReactJS, Nexus, and Rollup JS.

API Developer – International Consulting Services

Responsibilities included Working with API development team in creating and defining RESTful APIs. Leveraged AWS for hosting services, Spring Boot for API development, Jenkins for DevOps, and Swagger/Confluence for documentation.

Front End Developer – International Consulting Services

Responsibilities included leading frontend team in creating an experience and expertise search solution for the consulting firm to find experts in a reliable, repeatable and trustworthy experience. Worked closely with API development team in creating and defining RESTful APIs. Leveraged AWS for hosting services, and ReactJS for creating a single page application.

SKILLS

- Frontend Development
- Backend Development
- Cloud Strategy and Development
- Application Architecture
- Frontend frameworks (React, Vue, Angular)
- CMS platforms (Drupal, WordPress)
- API Development (Java, Node.js)
- Cloud Solutions (AWS, GCP)

CERTIFICATIONS

- AWS Cloud Practitioner

Neil Mager

Cloud Architect

Neil is an architect with over 10 years of experience spanning application, enterprise, and cloud architecture. With a software development background, he helps clients leverage technology to drive transformation and deliver meaningful business outcomes.

RELEVANT EXPERIENCE

Enterprise Architect – Education Technology
Developed a multi-year cloud optimization and adoption roadmap after the migration of 40+ applications. Defined strategies for containerization, CI/CD, and high availability while aligning stakeholders across security, operations, and software.

Cloud Architect – Education Technology
Designed a fully serverless AWS architecture for an API-based application managing 250M assets. Delivered an 8x performance improvement and reduced costs through AWS Lambda, API Gateway, and Fargate.

Cloud Architect – Large-Scale Migration
Partnered with application owners and AWS ProServe to plan and execute complex migrations of thousands of servers and hundreds of applications. Established repeatable migration processes and tracking systems to enable a scalable "migration factory" model.

SKILLS

- Current & Future State Architecture
- Business & Technical Analysis
- CI/CD & Value Stream Optimization
- Change Leadership
- Cloud Training
- Proof of Concept Design & Analysis
- Large-Scale Cloud Migration

CERTIFICATIONS

- AWS Cloud Practitioner
- AWS Associate Solutions Architect
- AWS Well-Architected Proficiency

Whitney Waite

Solution Owner – Scrum Master/Product Owner

Whitney is a technology leader with over 10 years of experience building high-performing teams and delivering business-critical platforms for federal and commercial clients. She specializes in product and project management, strategic planning, and large-scale program delivery.

RELEVANT EXPERIENCE

Solution Owner / Engagement Lead – Data Platform (Healthcare)
Led discovery and implementation of an HR data mart to enable analytics-driven HR strategy. Partnered with executives and SMEs to define approach, timeline, and Agile governance.

Senior Solution Owner – Enterprise Service Modernization & Data Transformation (Insurance / Non-Profit)
Directed development of a modern Azure-based data processing platform, including on-prem migration, data pipeline design, and new application build to support internal data management and auditing.

Senior Solution Owner – Application & Data Modernization (Government)
Led multi-team modernization of a large-scale legacy system from .NET to AWS. Established SRE practices, improved business processes, and managed Agile delivery, dependencies, and communications across programs.

Senior Solution Owner – Program (Government)
Implemented a quarterly planning cadence using modified SAFe for 10+ distributed product teams, aligning roadmaps, mitigating risks, and improving program transparency and delivery outcomes

SKILLS

- Strategic Planning & Analysis
- Agile/Scrum (SAFe Practitioner)
- Project & Product Management
- Requirements Gathering & Roadmapping
- Software & Web Development
- CloudFormation & Site Reliability Engineering
- Business & Data Analysis

Tamika Sterlin

Software Engineer

Tamika is a Software Engineer/Consultant with a background in Business Administration. She works as a full stack engineer with a focus on front-end development working with React. She enjoys learning new technologies and being a team player.

RELEVANT EXPERIENCE

Front-End Engineer – Gather App
Developing a mobile-friendly React web app enabling employees to create and join events, share images, and stay connected across the organization.

Front-End Engineer – Retail Toy & Gaming
Built a scalable full-stack e-commerce site using Chakra UI, PWA, React, and Salesforce. Created reusable React components and an engaging UI in collaboration with a cross-functional team.

Front-End Engineer – Healthcare
Delivered a multilingual, mobile-friendly React website for a large medical client, incorporating geolocation services and supporting translation into eight languages.

Front-End Engineer – Healthcare
Developed a Vue.js website for a major healthcare client based on Figma designs, creating interactive UI elements that showcased the client's partnership with Slalom.

Software Engineer – Financial Management
Implemented unit testing for Angular components, converted HTML logic into TypeScript, and participated in code reviews and application testing.

Front-End Engineer – Financial Management
Built UI components using React and Lit based on Figma designs for a financial services client. Created a reusable component library and supported backend REST API integration.

SKILLS

- React
- JavaScript
- PWA
- Vue.js
- Next.js
- Typescript
- Chakra UI
- Material UI
- React Native
- Figma
- Node.js
- Mulesoft
- Amplience
- Lit
- REST APIs
- Docker

Budget & Pricing (Attachment C)

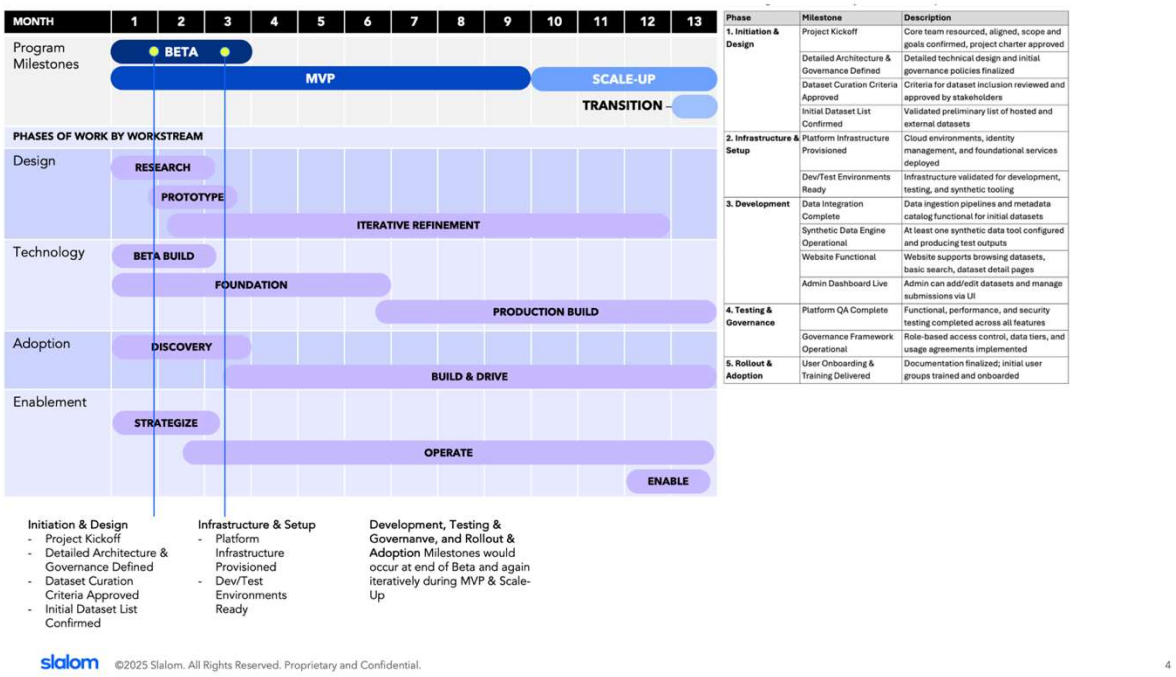
Please refer to the Attachment C - Services Budget Template for a breakdown of the costs associated with each phase of the proposed project.

Our plan is structured as a phased approach designed to deliver value to MassTech as early as possible. By sequencing development into distinct phases – Beta, MVP, and Scale-Up – we enable incremental releases that put working solutions in the hands of users within the first three months, with ongoing enhancements and incremental functionality throughout the project timeline. This strategy accelerates learning, supports rapid adoption, and prioritizes releases informed by real user feedback.

The overall estimated cost for delivering all phases and features described in our proposal is approximately $15 million. However, the phased release model means that initial launches – such as Beta and MVP phases – require only a portion of the total investment.

*Please note that Slalom is not an authorized reseller. The licensing costs included in this response are estimates based on information provided by our partners. Final pricing and licenses must be obtained directly from our partners or their authorized resellers.

Appendix – RFP Milestones



Phase	Milestone	Description
1. Initiation & Design	Project Kickoff	Core team resourced, aligned, scope and goals confirmed, project charter approved
	Detailed Architecture & Governance Defined	Detailed technical design and initial governance policies finalized
	Dataset Curation Criteria Approved	Criteria for dataset inclusion reviewed and approved by stakeholders
	Initial Dataset List Confirmed	Validated preliminary list of hosted and external datasets
2. Infrastructure & Setup	Platform Infrastructure Provisioned	Cloud environments, identity management, and foundational services deployed
	Dev/Test Environments Ready	Infrastructure validated for development, testing, and synthetic tooling
3. Development	Data Integration Complete	Data ingestion pipelines and metadata catalog functional for initial datasets
	Synthetic Data Engine Operational	At least one synthetic data tool configured and producing test outputs
	Website Functional	Website supports browsing datasets, basic search, dataset detail pages
	Admin Dashboard Live	Admin can add/edit datasets and manage submissions via UI
4. Testing & Governance	Platform QA Complete	Functional, performance, and security testing completed across all features
	Governance Framework Operational	Role-based access control, data tiers, and usage agreements implemented
5. Rollout & Adoption	User Onboarding & Training Delivered	Documentation finalized; initial user groups trained and onboarded

Figure 24: RFP Milestones

A mapping of our proposed project timeline against the milestones defined in the RFP.

Appendix – Customer References

Reference #1: Biogen

Contact Information:

Dave Clifford

Head of Data Science and Applied Machine Learning

Dave.Clifford@biogen.com

Services Delivered: Data Lakehouse, Custom Data Catalog, Data Platform Architecture Design, Cloud Based Solutions

Technologies: AWS, Databricks, Terraform, Kubeflow, 3D Slicer, MonAI

"The Slalom team has completed incredible work, which boils down to us onboarding new studies in a matter of weeks, and makes the data Findable, Accessible, Interoperable, and Reusable.

This is a big deal because it enables us and partner researchers and scientists to spend more of their time doing research and science (to help understand patient response therapies and develop new therapies that impact patients' lives), rat her than sourcing for what they need and asking a data scientist to pull them data. Simply put, it accelerates science and research." – *Dave Clifford, Head of Data Science and Applied Machine Learning*

Why

The Digital Health division of a leading neuroscience biotechnology company wanted to modernize the data platform that they use to manage their research, and participation in real world research networks in neurodegenerative diseases.

The client engaged Slalom to:

- Build a Unified Data Platform that would provide a flexible and scalable system to onboard, filter, and curate data of different varieties.
- Build an imaging pipeline for research and clinical informaticists to quality check DICOM images and annotate.
- Create a centralized Data Catalog of contents with study tags, data dictionaries, and lineage.

What

Following a six-week discovery phase, Slalom launched three agile delivery teams to develop across key domains 1) the foundational architecture of a new AWS and Databricks data platform, 2) Biobanking and Clinical Operations, and 3) imaging workflows. To streamline data onboarding, Slalom engineered configuration-driven ingestion frameworks

Reference #1: Biogen

capable of handling both structured and unstructured datasets from multiple sites. This enabled rapid integration of study data. Infrastructure-as-code was leveraged to optimize provisioning of Databricks workspaces, Spark clusters, and workflows. Collaborating closely with study owners, clinical informaticists, and researchers, the teams developed an automated consent-check solution that excludes unconsented data from downstream consumption layers and data products. In parallel, Slalom partnered with clinical sites to standardize Biobanking inventory data exchange templates and produce inventory reports. For imaging workflows, Kubeflow-based pipelines were built to support researchers in analyzing brain scans efficiently.

Wow

Slalom successfully onboarded a diverse range of datasets—including DICOM and HL7 v2 EMR—using YAML-driven configuration pipelines that enabled scalable and repeatable ingestion processes. To maintain data integrity, the team implemented automated quarantine protocols for records missing collection dates or consent forms. A comprehensive library of data quality rules was developed, allowing for automated validation checks through configurable workflows. These enhancements led to a sixfold improvement in study onboarding times. Additionally, Slalom delivered multiple data products, including curated data cuts that supported external study publications, and constructed a longitudinal patient view to help researchers better understand the full patient journey.

Reference #2: U.S. Citizenship and Immigration Services (USCIS)

Contact Information:

Denise Johnson

Denise.A.Johnson2@uscis.dhs.gov

202-236-3937

Services Delivered: Data, Artificial Intelligence, Cloud, Systems implementation, Experience strategy & design, Digital product building, Privacy & security

Technologies: AWS, Tableau, Databricks, React, Java Spring

Why

U.S. Citizenship and Immigration Services (USCIS), a component of the Department of Homeland Security, is one of the most progressive and customer-focused technology organizations in the federal government. Its verification systems are used by hundreds of thousands of people and processes millions of transactions annually. USCIS sought to enhance these systems to maintain reliability, improve the digital experience, and optimize costs. USCIS also needed to refresh its operational model, improve tooling and training, and consolidate its ecosystem from multiple siloed vendors into a single coordinated partner.

What

Over the course of 5 years, Slalom partnered with USCIS to modernize its cloud-based portfolio of 10 verification systems—including E-Verify and SAVE—through a multi-year engagement focused on reliability, performance, and cost optimization. The size of this project was $25 million per year, for the first three years.

Slalom consolidated four vendor contracts into one, implemented a modified scaled Agile framework, and deployed a 24/7 software support and operations model aligned to site reliability engineering (SRE) principles. Slalom introduced a culture of continuous innovation through hackathons, innovation sprints, and OKRs, and supported USCIS with assessment of team capability and maturity. Slalom led the transition from legacy systems, coordinated across seven products, and delivered annual product roadmaps to complete the full migration to AWS while continuing to enhance the features of each product.

Wow

Slalom delivered a modernized, cloud-native verification platform, improved USCIS's operational efficiency, and enhanced the digital experience of users. Key outcomes include:

- Reduced cloud infrastructure costs by 20%.
- Migrated from Oracle to AWS and implemented modern analytics, including Tableau and Databricks.
- Improved accessibility through 508-compliant development patterns.

Reference #3: Sarepta

Contact Information:

Bill Thronton

CIO, Sarepta Therapeutics

bthornton@sarepta.com

617-274-4000

Services Delivered: Enterprise data platform development, cloud system integration and modernization, data governance and regulatory compliance enablement, analytics performance optimization, and secure collaboration and workflow automation.

Technologies: AWS (Amazon Web Services), Snowflake, Informatica

Why: Sarepta's real-world evidence (RWE) team plays a critical role in ensuring treatments reach patients quickly, producing the data required for FDA and international regulatory submissions. But inside the organization, data lived in silos. With information spread across hundreds of systems, analysts needed 6–12 months just to prepare a single dataset for analysis. There was no enterprise-level infrastructure to bring everything together, and limited governance meant access was inconsistent and unclear.

In rare disease research, every data point matters. But with slow processes, disconnected systems, and reliance on external vendors, Sarepta's analysts couldn't act on insights quickly enough

What: Sarepta partnered with Slalom to design and implement a modern enterprise data platform (EDP) that would support both speed and scale. The platform would initially serve the RWE team but ultimately empower the entire organization.

Built on AWS with Snowflake and Informatica, the platform brings together Sarepta's clinical trial, claims, diagnostic, and observational data into a secure, governed environment. It was engineered for performance, flexibility, and regulatory compliance. Every dataset is traceable, validated, and approved for use.

Ensuring compliance was also critical. The EDP allows Sarepta to make the data standards required by global regulators, including the FDA. All requests are tracked, and sensitive datasets are protected under strict frameworks.

Wow The platform's impact was immediate, reducing data loading times from five days to two hours and expediting claims data integration from months to days. Data ingestion is now up to 99.8% faster, facilitating more patient participation in trials with transparent eligibility. Regulatory responses and internal research are significantly quicker, enabling the RWE team to support submissions confidently and swiftly.

Sarepta also transitioned to internal analysts, increasing agility and control. The data platform supports trial assessment and manufacturing, improving yields and reducing costs, while enhancing collaboration with a centralized data source. This cultural shift positions

Reference #3: Sarepta

data as a strategic asset, allowing faster decisions and helping Sarepta efficiently achieve its mission.

Appendix – Team Structure

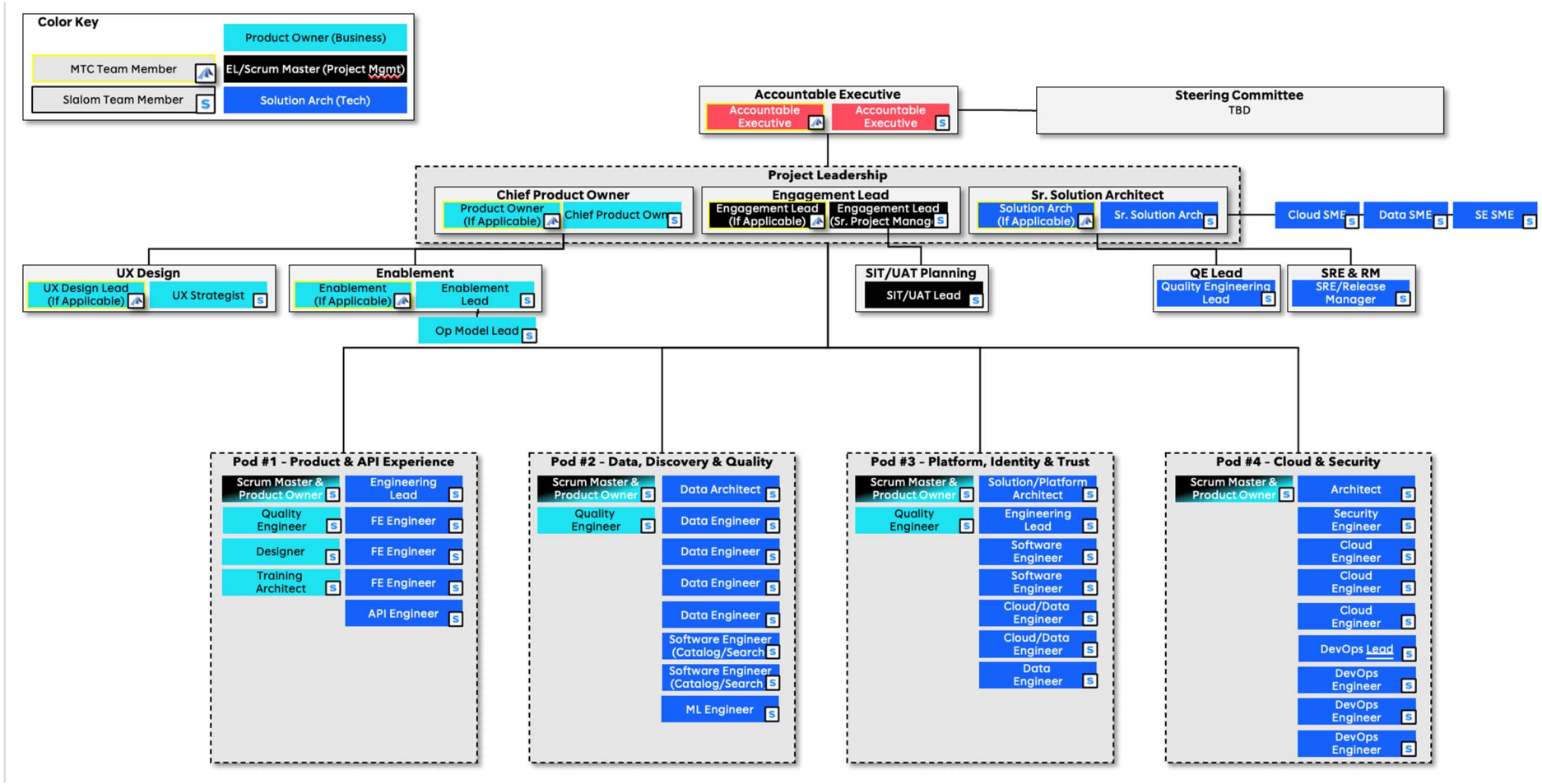


Figure 25: Team Structure

Appendix – Customer Stories

Within this section, we include selected stories that further illustrate Slalom's proven expertise in data warehouse modernization, responsible AI solutions, and large-scale technology implementations. While these examples are not formal references, they demonstrate the breadth and impact of our work across diverse industries and public sector engagements.

customer story: Banking

Data Exchange Unlocking Untapped Value in Banking

AWS | Snowflake

One of Australia's largest bank is increasingly partnering with businesses and institution to tap into one of Australia's largest aggregated and de-identified transaction banking datasets to analyse trends, overcome challenges, and create better, more targeted products and services for Australians.

They have a desire to develop a commercial and non-commercial cloud data exchange platform that is reusable safe and secure and can expedite data sharing initiatives for the bank and their partners to realise value at scale.

" being able to do this for business customers is where the real value is going to be, using collaborative analytics. Take our data, give it to them, they bring their data, start running ML

Client

One of Australia's largest s multinational bank with businesses across New Zealand, Asia, the United States and the United Kingdom

Solutions

Delivered Data Exchange operating model strategy to address gaps in process, standards, governance and support.

Delivered Data Exchange Architectural designs covering use case patterns in line with best practices to enable faster onboarding & deployment of future use cases.

Delivered feasibility assessment on Collaborative Analytics demonstrating how the solution can support data sharing, collaborative analysis with partners in a secure CBA environment.

Delivered Cyber Security reviews & Well Architecture Reviews to ensure the solution met cyber, risk and security protocols.

Platforms

AWS

Snowflake

Socializing data faster with an analytics marketplace

AWS Serverless | Tableau | Data Visualization | IT Spend Reduction | Streaming Television

Why

A global TV studio was seeking a partner to help build a reporting and data platform to serve as a consolidated repository where departmental data could be stored and subscribed to.

The new reporting platform led to increased user self-service and reporting flexibility. Additionally, it enabled better decision making and saved money and time by moving away from legacy reporting systems.

"The migration effort was massive, but the payoff is clear–fewer reports, faster access, and real savings." – Mark Shafer, SVP, Decision Science

What

Slalom worked with the Studio to deliver a cost effective, highly available, near real-time, on premise and cloud-based operational data and analytical data reporting solution that enables users to manage data security and access through the Client's Active Directory. The solution was designed as a hybrid reporting and data platform using AWS.

Slalom migrated data sources from two legacy reporting portals onto the platform.

Additionally, Slalom rationalized reporting systems and built out new reporting solutions with an emphasis on decreasing existing reports and leveraging data visualizations where a single data dashboard can replace several reports.

Wow

As a result of Slalom's partnership, the Client identified approximately 1,300 reports to migrate leading to:

- The elimination of 500 reports through rationalization and workflow operations
- Migration of 349 total reports for 57 key systems
- Addressing security concerns via an enterprise-wide cloud migration
- Reduced technology support costs by $435k annually
- Enabled business the unites the ability to self-service their reporting/data needs

Creating a modern culture of data in Life Sciences

Cloud Enablement | Data & Analytics | Azure | Snowflake | Power BI

Why

Following the commercial success of its flagship cancer treatment, and with two additional drugs in Phase 3 trials, this biopharmaceutical client needed to significantly modernize and scale its IT and data analytics infrastructure to be ready for rapid growth.

The company's leadership had decided the time was right to consider moving to cloud-native infrastructure and data platforms, but these capabilities were new to the organization.

What

Slalom first helped create a strategic roadmap for modernizing the company's cloud infrastructure and data capabilities.

Slalom helped the company select Microsoft Azure as its chosen cloud platform, and Snowflake and Power BI as its core data & analytics tools.

Over the following months, Slalom helped to implement the company's first-ever cloud platform by developing:

- Cloud infrastructure & security standards
- Azure resource management, monitoring, and tagging processes
- Snowflake cloud data warehouse
- Power BI training and enablement

Wow

In just a few short months, Slalom was able to help the company implement and operationalize a secure and scalable architecture in Microsoft Azure, create an enterprise data warehouse in Snowflake, and start delivering insights in Power BI.

By leveraging Slalom expertise and our agile approach, the client was able to:

- Establish a foundational blueprint for leveraging the cloud
- Provide access to data and insights in days or weeks, not months
- Begin decommissioning legacy on-premise hardware
- Explore and prototype big data and machine learning capabilities in Azure

Design & implement MDM for customer accounts across NA to establish data consistency across systems & eliminate duplicate accounts for better & consistent reporting & analytics.

Industry: Healthcare | **Domain:** Master Data Management | **Technology:** Boomi MDM, Veeva CRM & Snowflake

why

With the increase in customer base of Seattle Genetics, the client could not focus on taking steps to ensure that data in its primary customer systems – CRM is complete, clean, standardized & devoid of duplicates. As a result, client was facing the following issues:

a. Inconsistent customer data across systems leading to confusion & slow decision making of business decisions.

b. Duplicate customer accounts in CRM limiting their ability to analyze data at customer level..

c. Data quality issues in customer account attributes were starting to surface with no way to identify or eliminate them to downstream systems.

d. Lastly, utility of reporting & analytics layer without trustworthy data would have been very limited.

what

Designed & implemented master data management hub using Boomi MDH module. The tool was integrated with CRM as its source and Snowflake as data warehouse which was the gate for all reporting & analytics requirements. Specifically -

a. Implemented Boomi master data hub module to master all NA (US + CA) customer accounts to build repository of clean, standard & unique accounts.

b. Integrated Boomi MDH with Snowflake to ensure trustworthy customer master data is available for downstream applications and/or reporting purposes.

c. Ensured the process to identify accounts with data quality and/or duplicate issues works in future for new accounts as well.

d. Ensured architecture adopted allows implementation of MDM for customers in new geographies (EU) as well.

wow

Ability to have single data repository with all customer accounts, irrespective of sources, and holding trustworthy data with all agreed pre-defined data attributes was the biggest win. Specifically-

a. All users of Snowflake data warehouse, now, see one version of customer master data rather than N versions.

b. Reporting across systems becomes consistent as they are now looking at single source of truth.

c. Business can take better & faster decisions based on reporting as reports can now be trusted due to clean & consistent data as its source.

d. Architecture allows more geographies to add their respective customers to same Boomi MDM instance to allow reporting across geographies.

slalom | SAREPTA THERAPEUTICS

Learn more about the why behind this work, what we did, and how it's driving impact

> "By doing this work, we're allowing our data scientists to be scientists. There's a **tremendous efficiency and innovation** that comes when you give them back time that was spent searching for and gathering data."
>
> Bill Thornton
> CIO, Sarepta Therapeutics

Why

Sarepta's real-world evidence (RWE) team plays a critical role in ensuring treatments reach patients quickly, producing the data required for FDA and international regulatory submissions. But inside the organization, data lived in silos. With information spread across hundreds of systems, analysts needed 6–12 months just to prepare a single dataset for analysis. There was no enterprise-level infrastructure to bring everything together, and limited governance meant access was inconsistent and unclear.

In rare disease research, every data point matters. But with slow processes, disconnected systems, and reliance on external vendors, Sarepta's analysts couldn't act on insights quickly enough.

What

Sarepta partnered with Slalom to design and implement a modern enterprise data platform (EDP) that would support both speed and scale. The platform would initially serve the RWE team but ultimately empower the entire organization.

Built on AWS with Snowflake and Informatica, the platform brings together Sarepta's clinical trial, claims, diagnostic, and observational data into a secure, governed environment. It was engineered for performance, flexibility, and regulatory compliance. Every dataset is traceable, validated, and approved for use.

Ensuring compliance was also critical. The EDP allows Sarepta to meet the data standards required by global regulators, including the FDA. All requests are tracked, and sensitive datasets are protected under strict frameworks.

Wow

The platform's impact was immediate, reducing data loading times from five days to two hours and expediting claims data integration from months to days. Data ingestion is now up to 99.8% faster, facilitating more patient participation in trials with transparent eligibility. Regulatory responses and internal research are significantly quicker, enabling the RWE team to support submissions confidently and swiftly.

Sarepta also transitioned to internal analysts, increasing agility and control. The data platform supports trial assessment and manufacturing, improving yields and reducing costs, while enhancing collaboration with a centralized data source. This cultural shift positions data as a strategic asset, allowing faster decisions and helping Sarepta efficiently achieve its mission.

BROADBAND FIRM DATA CATALOG

Building a Modern Culture of Data Using Alation

Modern Culture of Data | Data Catalog | Alation | MVP Implementation

Why

A broadband connectivity company grew rapidly and now dominates the market for broadband and connectivity services. The company now has millions of customers across 41 states. Despite the company's growth, information at the company remained siloed making it difficult to make enterprise-wide decisions.

To scale in the future and continue to flourish, it wanted to harness the power of its siloed data ecosystem and empower its employees to trust and use data for decision making.

What

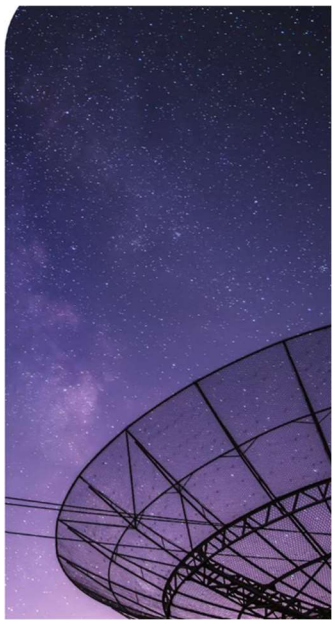
Our approach was to enable a Modern Culture of Data centered around a data catalog implementation. We implemented intuitive, single solutions for end data consumers built on best-in-class tools and integrated governance.

By choosing a single technology, Alation, we were able to provide visibility to siloed data, recommend opportunities to reduce workload redundancies, and provide visibility to better-quality data.

Wow

By connecting 30 disparate data silos, adopting a new lightweight data governance process, and embracing a balance of data certification and self-service, the Fortune 100 broadband firm empowered hundreds of employees to make better decisions based on data.

Today, 45 departments and more than 200 users count on the solution to support critical daily decisions, from knowing where congestion is occurring in the network, to reclaiming IP addresses, to identifying network service theft.



slalom | UNITED

Generative AI lifts traveler experience

With a strong culture of innovation, United Airlines was eager to implement generative AI (GenAI). Together, we developed solutions that empower employees and elevate the customer experience.

In less than 60 days, we built a platform that securely allows United's teams to experiment with GenAI. After experimenting with and evaluating more than 30 use cases, United developed two solutions with Slalom:

- New capabilities for increasing customized messages about flight status changes
- A customer care copilot that helps agents provide faster and more empathetic responses to customer inquiries

The resulting efficiency and transparency from the new solutions have strengthened customer experience and trust.

Impact

PEOPLE PLANET PERFORMANCE

30
GenAI use cases tested and evaluated in less than 60 days

35%
More flight updates are customized for travelers thanks to GenAI

2x
Faster mean time to resolution with the customer care copilot

Customer

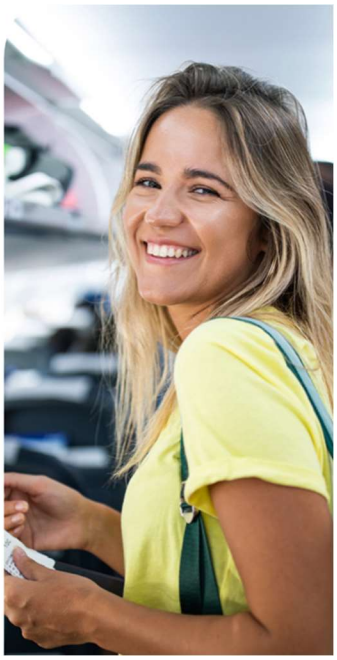
As the world's largest airline, United Airlines has an industry-leading loyalty program, an award-winning app loved by travelers, and routes to more than 210 destinations in the US and more than 120 internationally.

Services

Strategy
Data
Artificial intelligence
Systems implementation
Experience strategy & design
Digital product building
Operations
Privacy & security

Partners

Amazon Web Services (AWS)
Anthropic (Claude)



slalom | NORWEGIAN CRUISE LINE

Developing a GenAI assistant to help guests plan their dream cruise vacation

A guest can make up to 64 considerations before booking a cruise. Recognizing this complex digital journey, Norwegian Cruise Line (NCL) partnered with Slalom to simplify the booking experience by leveraging generative AI (GenAI). Together, we launched Nora, a GenAI-driven search assistant that narrows down itinerary options so guests can more easily plan their dream cruise vacations.

Powered by OpenAI and AWS, our Nora uses conversational language to tailor cruise recommendations and answer questions based on user inputs. By streamlining vacation planning for guests, this GenAI assistant is helping NCL drive more bookings, at a higher value.

Impact

PEOPLE PLANET PERFORMANCE

94.5%
Of user sentiment is positive during chat conversations with Nora

58%
More likely for guests to enter the booking funnel after interacting with Nora

+16%
Increase in the value of bookings from guests who interact with Nora

Customer

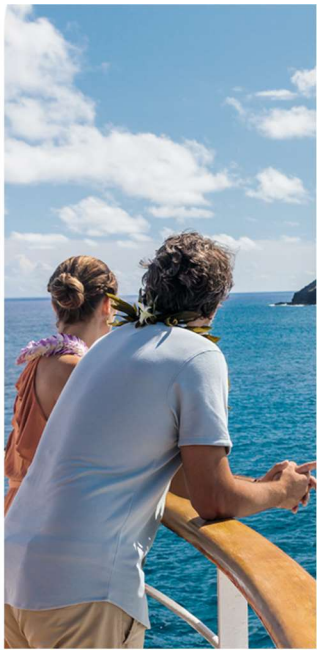
Norwegian Cruise Line is one of the largest cruise lines in the world, with an award-winning fleet of ships and over 400 travel destinations.

Solutions

Generative AI
Experience strategy & design
Data & technology strategy
Customer & employee experience
Product strategy, design & innovation

Partners

OpenAI
Amazon Web Services (AWS)
LangGraph



COVID Transparency for 30+ Million Californians

To defeat the spread of COVID-19, the California Department of Technology (CDT) began sourcing data from multiple state agencies and local partners to support data-driven decision making for the Office of the Governor. By late March, the CDT needed support consolidating, standardizing, and visualizing pandemic data. Enter Slalom.

Using Snowflake as the centralized repository for state and local data sources, Slalom developed both executive and public facing dashboards comprising case statistics, testing, hospitalizations, equipment, supplies, and many more key data points.

Live within days, Slalom's dashboards rapidly enabled the governor's office and public health officials to make more effective and better-informed decisions.

6→1

Dashboards consolidated

+1 mil

Public dashboard views in first month

72

Hours required to design, build, and deploy the dashboards

Client

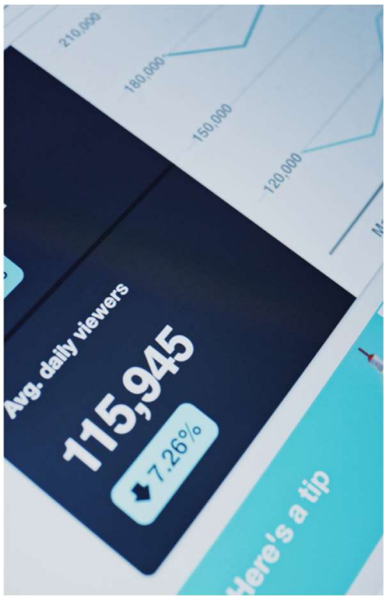
The California Department of Technology (CDT) organizes and presents critical data supporting informed decisions by the Governor to protect the health and safety of all Californians.

Solutions

Snowflake + Tableau
Dashboarding
Web design
Data Ingestion and Engineering

Platforms

Tableau
Snowflake



A data platform to handle anything.

A healthcare information exchange wanted to analyze its data faster and curate deeper insights. Together, Slalom and CRISP built an Azure Data Lake. The results forever changed how CRISP serves its stakeholders—and its local community now benefits from some of the most reliable, insightful healthcare information in the US.

Why

CRISP, a healthcare information exchange, wanted to analyze its data faster and curate deeper insights. CRISP and Slalom needed to automate and re-engineer the current processes that reported vital COVID-19 patient data to the state to perform contact tracing, and so its citizens could remain informed.

What

Slalom was able to build the cloud infrastructure in about three weeks using our Azure Data Platform accelerator. Concurrently, our data engineers worked with the healthcare experts at CRISP to transfer their business logic for combining and standardizing many different data sources across different companies into curated outputs for the state government and contact tracers to use.

Wo w

Through our partnership with CRISP, we were able to produce and deliver hourly contact tracing datasets to the state's contact tracers, allowing them to perform their jobs with timely information. Additionally, Slalom was able to perform the original "pre-COVID" scope, enabling the CRISP team to have an extensible analytics environment that allows them to add new data sources to the platform in less than an hour.

Solutions

Modern Data Architecture
Data Engineering
DataOps

Technologies

Azure
Databricks

> **We went from this 12-hour day, multiple-person process that we had to do every single day, including weekends, to an automated file that you spit out in a couple of hours. - *Lindsey Ferris,* Program Director, CRISP**

Databricks Platform Modernization & Lakehouse Architecture

Why

Nasdaq & Slalom set out to standardize their index data into a platform to provide a single source of truth across business intelligence, data science, feed machine learning models and other applications.

All their legacy data platforms and frameworks were tied to complex architectures that were inhibiting their ability to meet rapid time to mark.

What

Slalom & Nasdaq identified data sets to bring onto the Databricks Lakehouse platform, launched two teams focused on data and the API layers, and **built a scalable, repeatable cloud architecture**.

Slalom brought our expertise around scalability, reliability and repeatability and embedded our team with Nasdaq's to create repeatable patterns & playbooks for future datasets to be migrated onto the platform.

Wow

Nasdaq is building the platform to be the cornerstone of their data strategy with Index Technology. The new platform provides a single trusted data source for multi-use across the enterprise.

Slalom streamlined scalable data with history and data was now managed and designed for recoverability. Additionally, we built a playbook to drive adoption.

Unified Data Layer with Databricks & AWS

Why

A leading global financial services organization's Cybersecurity & Intelligence (C&I) team partnered with Slalom and Databricks on a pilot that aimed to reimagine and build key data processing and business insight use-cases using Databricks on AWS.

This organization needed a solution that allowed downstream customers readily available and accurate data & insights, while being highly performant, cost-effective, and manageable at scale.

What

Slalom was tasked with **developing out a production ready infrastructure** that could handle production-like data unification, while understanding best patters to achieving best performance & cost.

Key deliverables for this solution were:

- **Infrastructure Readiness**: Standing up an MVP Databricks environment on AWS
- **Unified Data Layer Architecture**: Develop architecture & execute a production-like process for UDL
- **Performance & Cost**: Understand performance & cost ramifications.
- **Best Practices & Knowledge Transfer**: Established repeatable patterns around Databricks & architecture around UDL

Wow

Also, consider: Is there a bigger picture in terms of the change we enabled for the client, why this story would be interesting to other clients/prospects, or how it specifically showcases our brand/strengths?

Consider using bullets to highlight:

- Fully compliant with AWS cloud security provisions
- 20 + users using the Platform running fully managed pipelines
- Creation billions of structured synthetic data that can be scaled as needed
- Developed 6 streaming pipelines to demonstrate unification data layer
- Created repeatable patterns, developed decision logs and maintained collection of best practices to be used for the future.

Solutions

Cloud infrastructure, unified data layer architecture, performance & cost assessments, best practices & knowledge transfer

Technologies

Databricks

AWS

Cybersecurity

PEM/Agile

Modern Data Analytics Platform

Big Data | Data Visualization | Training | Program Management | Analytics | Production Support

Background

A large, Canadian Credit Union was looking to develop a modern information management and analytics system with the technology, people capabilities and processes to protect, manage and leverage their data assets to support evidence-based decision making, innovation with ML/AI and growth of the organization.
Having already selected Azure as their new Cloud platform, this organization was 5 months into an in-house implementation that was yielding little results when they reached out for assistance and expertise.

Project

Slalom landed a small pod (3 times smaller than the credit union's team) of cloud experts who proceeded to review the organizations target objectives, jointly create a new target architecture and select a number of high priority use cases to implement was part of a MVP solution.
Over the course of 8 weeks, the Slalom team quickly created the patterns to ingest data, create a data lake, model data into a warehouse and development a set of custom dashboards in PowerBI effectively completing more scope in 2 months compared to the original 5.

Results

The success of the MVP laid the foundation elements of the modern information management and analytics platform for this credit union.
The slalom team created a **dynamic data ingestion accelerator** as part of the next phase thereby reducing the time to on-board a new data source **from 4-6 weeks down to 1-2 weeks**. Using this, the Azure data platform has quickly grown to meet and exceed the many use cases required by the credit union.

Transforming AI Governance in Public Safety

AI Governance | ISO Alignment | Public Safety

Why

A leader in public safety technology recognized an urgent need to elevate its AI governance framework in response to an evolving regulatory landscape and growing customer inquiries regarding Responsible AI practices. The discovery phase illuminated critical gaps in their decentralized governance model, revealing opportunities to unify and streamline AI operations. Our mission was to architect a robust governance structure and operating model that not only aligns with ISO standards but also empowers the safety innovation company to scale its business with confidence and integrity, ultimately ensuring public trust and safety.

What

To achieve this transformation, we embarked on a strategic delivery encompassing:

- A roadmap to transition from a decentralized to a federated AI governance model that enhances role clarity and stakeholder coordination.
- Recommendations aligned to relevant ISO control objectives to expedite ISO certification.
- Strengthening AI impact assessments and mapping them to the AI system lifecycle for improved accountability.
- Leveraging existing policies and procedures to establish a single source of truth for AI governance.
- Harmonizing data classification and sensitivity across the organization to ensure compliance and privacy.

Wow

The project culminated in remarkable outcomes that underscored our expertise and commitment:

- Established a federated governance model that reduced operational friction and increased decision-making velocity.
- Enhanced role clarity and streamlined processes, accelerating the path to ISO certification by 30%.
- Developed a robust AI system inventory, improving transparency and readiness for potential audits and customer inquiries.
- Fostered an environment of innovation that balanced regulatory compliance with cutting-edge technology deployment.
- Positioned client as an industry leader in Responsible AI, significantly boosting stakeholder confidence and market reputation.

Governing AI for a Global Reinsurance Leader

AI Governance | Privacy | Consumer Protection | Strategy

Why

A Fortune 500 reinsurance company, with nearly $100 billion in assets and over $3 trillion life insurance in force, sought to develop a comprehensive strategy to govern its increasingly widespread AI technology and solutions.

As a global leader, the company needed a strategy to establish processes, procedures, and structures for AI governance, ensuring legal compliance across its diverse operational regions, including North and South America, EMEA, and APAC.

Seeking subject matter expertise, the reinsurance company engaged Slalom.

What

Partnering closely with the global reinsurance leader, the Slalom team initiated the work with a maturity assessment followed by series of workshops that led to developing a strategy encompassing the following:

- AI Risk Management process
- AI Governance Structure including the formation of a program office and oversight board
- AI Systems Intake Process
- Developing Responsible AI principles
- Defining AI, a previously contested issue
- Recommendations to support the creation of AI Governance policies
- Model Governance Guidelines

Wow

Slalom delivered a comprehensive 6-month roadmap to the client, outlining processes, tools, and the necessary framework to operationalize AI Governance and ensure the responsible development and deployment of AI systems.

A key stakeholder remarked, *"Slalom has given us a significant head start"* through the deliverables produced from this collaborative effort.

This work will lead to a robust AI Risk Management Process, creation of a Responsible AI Program Office, and the alignment to key industry best practices like:

1. ISO/IEC 42001:2023 Information Technology–Artificial Intelligence–Management System
2. NIST AI 100-1 AI Risk Management Framework

ML Based Quality Assurance for Global Auto Manufacturer

A global automotive manufacturing company leaders partnered with Slalom to develop a solution for **synthetic data generation**. The synthetic data was designed to assist in the training of AI models for **automated quality assurance** in a manufacturing environment.

The client needed an automated method for detecting quality issues in flawed parts and components. Slalom created a tool to generate 3D models of parts in near real time.

The client's goal is 0 defects in manufacturing parts.

Connecting client with technology solutions

Providing infrastructure and data pipelines

Solving manufacturing challenges

Customer

Slalom helped the Global Automotive client connect with technology providers like AWS and NVIDIA to explore new possibilities and modern solutions.

Solutions

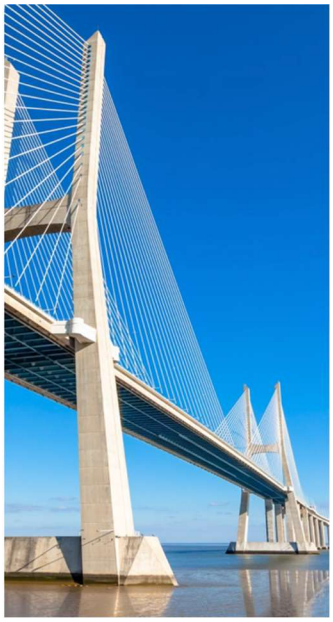
Converting 2D images to a 3D Neural Graphics Primitives (NGP) model in real-time

Creating a pipeline to add variations on existing CAD files, synthetically creating flaws, different materials, and different lighting conditions

Building infrastructure on AWS to run NVIDIA's Omniverse Replicator

Feedback

"There are a lot of problems we can solve with this technology" – Automation Engineering Manager



Detecting Wildlife with Computer Vision Models and Transfer Learning

A state department of transportation (DOT) is responsible for safe roads, which means understanding driving hazards, including those posed by wildlife. The DOT needed an automated solution to identify and count wildlife in images from remotely operated cameras.

Slalom partnered with the client to review their existing labeled data from cameras located at wildlife corridors, curate a larger dataset using images from similar projects and synthetic data generation and implement an object detection model with bounding boxes.

Together, Slalom and the client created a solution to reduce wildlife image analysis time. The resulting estimates of wildlife near roads will be used in future data-driven decisions.

Customer

The Department of Transportation exists to ensure that the state has a safe and efficient highway system by building and maintaining interstates, U.S. highways and state highways.

Performance

Specialized object detection in a variety of natural and infrared lighting conditions

Solutions

Multi-class object detection identified humans, animals, and vehicles.

Slalom curated the dataset used for Machine Learning model training.

Technology

Cloud Provider: GCP

Object Detection Model: You Only Look Once (YOLO) algorithm

slalom

VERIPAD

Detecting fraudulent medication with machine learning

Machine learning | Google Cloud enablement | Research innovation | Community

> Falsification of medicines is a big problem in many developing countries, where technological infrastructure is inadequate to detect these harmful products. Using Paper Analytical Devices (PADs) and machine learning, we can now quickly and efficiently classify drugs.
>
> Jason Ki
> CTO
> Veripad

BACKGROUND

The World Health Organization estimates that one in ten medical products in low- and middle-income countries is either substandard or falsified. That means vital medicines—like antibiotics, pain relievers, and anti-malarial drugs—are often worthless and sometimes deadly. PADs are the only product on the market that enables quick and easy chemical quality assessments for drugs on-site. A PAD is a simple paper card that carries several chemical tests and allows for the identification of the components of various medications. Veripad developed a mobile app that makes use of computer vision and machine learning to enable rapid, automated analysis of PADs—even in remote locations.

PROJECT

Slalom data scientists partnered with Veripad on a proof-of-concept to use machine learning to improve its PAD reading accuracy. We used Amazon SageMaker to quickly build the machine learning environment.

We worked with Veripad to apply image processing techniques like noise cancellation and contrast change to improve the quality of each image. From there, we used clustering techniques to differentiate color patterns and finally logistic regression to classify the type of drug and also assess its purity.

RESULTS

In just five weeks, we helped drive Veripad's fraudulent medication detection system to an accuracy above 90%, 5-10% higher than most human experts. This is a significant step to building user confidence in the app. "Our short time Slalom laid a foundation for us to continue working," says Ki. "We want to get close to 100% accuracy."

The team hopes one day Veripad will be used worldwide—not only to detect counterfeit medications but to gather data that tracks their spread, identifies patterns, and saves even more lives.

Form **W-9**
(Rev. March 2024)
Department of the Treasury
Internal Revenue Service

Request for Taxpayer Identification Number and Certification

Go to *www.irs.gov/FormW9* for instructions and the latest information.

Give form to the requester. Do not send to the IRS.

Before you begin. For guidance related to the purpose of Form W-9, see *Purpose of Form*, below.

Print or type. See ***Specific Instructions*** on page 3.

1 Name of entity/individual. An entry is required. (For a sole proprietor or disregarded entity, enter the owner's name on line 1, and enter the business/disregarded entity's name on line 2.)

Slalom, Inc.

2 Business name/disregarded entity name, if different from above.

3a Check the appropriate box for federal tax classification of the entity/individual whose name is entered on line 1. Check only **one** of the following seven boxes.

☐ Individual/sole proprietor ☑ C corporation ☐ S corporation ☐ Partnership ☐ Trust/estate

☐ LLC. Enter the tax classification (C = C corporation, S = S corporation, P = Partnership) . . . .

Note: Check the "LLC" box above and, in the entry space, enter the appropriate code (C, S, or P) for the tax classification of the LLC, unless it is a disregarded entity. A disregarded entity should instead check the appropriate box for the tax classification of its owner.

☐ Other (see instructions)

4 Exemptions (codes apply only to certain entities, not individuals; see instructions on page 3):

Exempt payee code (if any) 5

Exemption from Foreign Account Tax Compliance Act (FATCA) reporting code (if any)

(Applies to accounts maintained outside the United States.)

3b If on line 3a you checked "Partnership" or "Trust/estate," or checked "LLC" and entered "P" as its tax classification, and you are providing this form to a partnership, trust, or estate in which you have an ownership interest, check this box if you have any foreign partners, owners, or beneficiaries. See instructions . . . . . ☐

5 Address (number, street, and apt. or suite no.). See instructions.

255 S. King St., Suite 1800

Requester's name and address (optional)

6 City, state, and ZIP code

Seattle, WA 98104

7 List account number(s) here (optional)

Part I Taxpayer Identification Number (TIN)

Enter your TIN in the appropriate box. The TIN provided must match the name given on line 1 to avoid backup withholding. For individuals, this is generally your social security number (SSN). However, for a resident alien, sole proprietor, or disregarded entity, see the instructions for Part I, later. For other entities, it is your employer identification number (EIN). If you do not have a number, see *How to get a TIN*, later.

Note: If the account is in more than one name, see the instructions for line 1. See also *What Name and Number To Give the Requester* for guidelines on whose number to enter.

Social security number

			–			–				

or

Employer identification number

8	4	–	1	2	4	6	8	8	7

Part II Certification

Under penalties of perjury, I certify that:

1. The number shown on this form is my correct taxpayer identification number (or I am waiting for a number to be issued to me); and
2. I am not subject to backup withholding because (a) I am exempt from backup withholding, or (b) I have not been notified by the Internal Revenue Service (IRS) that I am subject to backup withholding as a result of a failure to report all interest or dividends, or (c) the IRS has notified me that I am no longer subject to backup withholding; and
3. I am a U.S. citizen or other U.S. person (defined below); and
4. The FATCA code(s) entered on this form (if any) indicating that I am exempt from FATCA reporting is correct.

Certification instructions. You must cross out item 2 above if you have been notified by the IRS that you are currently subject to backup withholding because you have failed to report all interest and dividends on your tax return. For real estate transactions, item 2 does not apply. For mortgage interest paid, acquisition or abandonment of secured property, cancellation of debt, contributions to an individual retirement arrangement (IRA), and, generally, payments other than interest and dividends, you are not required to sign the certification, but you must provide your correct TIN. See the instructions for Part II, later.

Sign Here | **Signature of U.S. person** [signature] **Date** 1/9/2025

General Instructions

Section references are to the Internal Revenue Code unless otherwise noted.

Future developments. For the latest information about developments related to Form W-9 and its instructions, such as legislation enacted after they were published, go to *www.irs.gov/FormW9*.

What's New

Line 3a has been modified to clarify how a disregarded entity completes this line. An LLC that is a disregarded entity should check the appropriate box for the tax classification of its owner. Otherwise, it should check the "LLC" box and enter its appropriate tax classification.

New line 3b has been added to this form. A flow-through entity is required to complete this line to indicate that it has direct or indirect foreign partners, owners, or beneficiaries when it provides the Form W-9 to another flow-through entity in which it has an ownership interest. This change is intended to provide a flow-through entity with information regarding the status of its indirect foreign partners, owners, or beneficiaries, so that it can satisfy any applicable reporting requirements. For example, a partnership that has any indirect foreign partners may be required to complete Schedules K-2 and K-3. See the Partnership Instructions for Schedules K-2 and K-3 (Form 1065).

Purpose of Form

An individual or entity (Form W-9 requester) who is required to file an information return with the IRS is giving you this form because they

About Slalom

Slalom is a purpose-led, global business and technology consulting company. From strategy to implementation, our approach is fiercely human. In 12 countries and 52 markets we deeply understand our customers—and their customers—to deliver practical, end-to-end solutions that drive meaningful impact. Backed by close partnerships with over 700 leading technology providers, our 10,500+ strong team helps people and organizations dream bigger, move faster, and build better tomorrows for all. We're honored to be consistently recognized as a great place to work, including being one of Fortune's 100 Best Companies to Work For seven years running. Learn more at slalom.com.

© 2025 Slalom, Inc. All rights reserved.

slalom

slalom.com

Massachusetts Technology Collaborative

Grant Agreement
Standard Budget and Invoice Template
Invoiced Costs

Line Number	Budgeting Applicant Information																			
2																				
3	Applicant: Slalom, Inc.				Procurement No.: 2026-JAII-01															
4	Budget Period: December 2025 - December 2026				Title of Proposed Project: AI Hub Data															
5	Address:																			
6	399 Boylston St, STE 1000																			
7	Boston, MA 02116																			
8					MTC Funding Requested:															
9					Total Project Cost:															
10	Cost Elements				Total Project	Invoice 1		Invoice 2		Invoice 3		Invoice 4		Cumulative						
11						Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY						
12							Amount		Amount		Amount		Amount	Expended		Remaining				
13	I. Direct Labor																			
14	*name/title*	*HRS*		*Rate*		*HRS*		*HRS*		*HRS*		*HRS*								
15	Latin America - Associate - Engineer	2520.0	$	87.00	$219,240		$ -		$ -		$ -		$ -	$ -		$	219,240.00			
16	Latin America - Junior - Engineer	11320.0	$	102.00	$1,154,640															
17	Latin America - Intermediate - Engineer	3880.0	$	128.00	$496,640															
18	Latin America - Proficient - Engineer	2280.0	$	153.00	$348,840															
19	United States - Associate - Engineer	1520.0	$	220.00	$334,400															
20	United States - Junior - Engineer	10040.0	$	235.00	$2,359,400															
21	United States - Junior - Product Owner/Scrum Master	2120.0	$	235.00	$498,200															
22	United States - Intermediate - Engineer	2720.0	$	255.00	$693,600															
23	United States - Intermediate - Product Owner/Scrum Master	1040.0	$	255.00	$265,200															
24	United States - Intermediate - Learning Architect	720.0	$	255.00	$183,600															
25	United States - Intermediate - Experience Designer	520.0	$	255.00	$132,600															
26	United States - Proficient - Product Owner/Scrum Master	4240.0	$	265.00	$1,123,600															
27	United States - Proficient - Operating Model	320.0	$	265.00	$84,800		$ -		$ -		$ -		$ -	$ -		$	84,800.00			
28	United States - Proficient - Engineer	3880.0	$	265.00	$1,028,200		$ -		$ -		$ -		$ -	$ -		$	1,028,200.00			
29	United States - Proficient - Experience Design	1440.0	$	265.00	$381,600		$ -		$ -		$ -		$ -	$ -		$	381,600.00			
30	United States - Proficient - Architect	3520.0	$	265.00	$932,800															
31	United States - Proficient - DevOps/Release Manager	2720.0	$	265.00	$720,800															
32	United States - Senior - Product Owner	2120.0	$	291.00	$616,920															
33	United States - Senior - SME	5574.0	$	291.00	$1,622,034															
34	United States - Senior - Engineer	1000.0	$	291.00	$291,000															
35	United States - Lead - Solution Lead	2200.0	$	291.00	$640,200															
36	United States - Executive - Project Executive	2550.0	$	321.00	$818,550		$ -		$ -		$ -		$ -	$ -		$	818,550.00			
37							$ -		$ -		$ -		$ -	$ -		$	-			
38	Total Direct Labor				$14,946,864	0	$ -	0	$ -	0	$ -	0	$ -	$ -		$	2,532,390.00			
39																				
40	II. Direct Labor Fringe Cost	Rate (%):			$0		$ -		$ -		$ -		$ -	$ -		$	-			
41																				
42	II. Subcontractors/Consultants/Contracted Staffing																			
43	*Firm name/title or role/type of services to be provided*	*HRS*		*Rate*		*HRS*		*HRS*		*HRS*		*HRS*								
44					$0		$ -		$ -		$ -		$ -	$ -		$	-			
45					$0		$ -		$ -		$ -		$ -	$ -		$	-			
46					$0		$ -		$ -		$ -		$ -	$ -		$	-			
47					$0		$ -		$ -		$ -		$ -	$ -		$	-			
48					$0		$ -		$ -		$ -		$ -	$ -		$	-			
49					$0		$ -		$ -		$ -		$ -	$ -		$	-			
50	Total Subcontractors/Consultants				$0	0	$ -	0	$ -	0	$ -	0	$ -	$ -		$	-			
51																				
52	III. Direct Materials (provide description and costs basis)																			
53	User Research Tool (i.e., Userlytics, User Interviews)				$5,000											$0		#REF!		
54	AWS - Estimated license costs				$17,129											$0		$5,000		
55	Snowflake - Estimated license costs				$17,282											$0		$17,282		
56	Databricks - Estimated license costs				$2,520											$0		$2,520		
57					$0											$0		$0		
58					$0											$0		$0		
59					$0											$0		$0		
60	Total Direct Materials				$41,930		$0		$0		$0		$0	$0			#REF!			
61																				
62	IV. Travel															$0		$0		
63																				
64	V. Other Direct Costs *(list type and description such as venue rental costs and cost basis)*																			
65					$0											$	-	$	-	
66					$0											$	-	$	-	
67					$0											$	-	$	-	
68					$0											$	-	$	-	
69					$0											$	-	$	-	
70	Total Other Direct Costs				$0		$ -		$ -		$ -		$ -	$ -		$	-			
71																				
72	VI. General & Administrative Expense/Overhead	Rate (%):			$0		$ -		$0		$0		$0	$ -			$0			
73																				
74	**Total PROJECT COSTS**				**$14,988,794**		**$0**		**$0**		**$0**		**$0**	**$0**			**#REF!**			

LEGEND:
Formula cell
Data entry cell

This section is for post-contract invoicing only. Do not complete with proposal.